

## Full paper

# Chemometric modeling of power conversion efficiency of organic dyes in dye sensitized solar cells for the future renewable energy

Jillella Gopala Krishna<sup>a</sup>, Probir Kumar Ojha<sup>b,1</sup>, Supratik Kar<sup>c,1</sup>, Kunal Roy<sup>b,\*\*</sup>, Jerzy Leszczynski<sup>c,\*</sup>

<sup>a</sup> Department of Pharmacoinformatics, National Institute of Pharmaceutical Educational and Research (NIPER), Chunilal Bhawan, 168, Maniktala Main Road, 700054, Kolkata, India

<sup>b</sup> Drug Theoretics and Cheminformatics Laboratory, Department of Pharmaceutical Technology, Jadavpur University, 188 Raja S C Mullick Road, 700032, Kolkata, India

<sup>c</sup> Interdisciplinary Center for Nanotoxicity, Department of Chemistry, Physics and Atmospheric Sciences, Jackson State University, Jackson, MS, 39217, USA



## ARTICLE INFO

## Keywords:

Chemometrics  
DSSC  
Dye  
PCE  
QSPR  
Solar cell

## ABSTRACT

The present study reports chemometric modeling of power conversion efficiency (PCE) of dye sensitized solar cells (DSSCs) using the biggest available data set till date which comprises around 1200 dyes covering 7 chemical classes. To extract the best structural features required for higher PCE, we have developed multiple partial least squares (PLS) quantitative structure-property relationship (QSPR) models for the Triphenylamine, Phenothiazine, Indoline, Porphyrin, Coumarin, Carbazole and Diphenylamine chemical classes using descriptors derived from the best subset selection method followed by selection of best five models in each dataset based on the Mean Absolute Error (MAE) values. The models were validated both internally and externally followed by the consensus predictions employing “Intelligent Consensus Predictor” tool to examine whether the quality of predictions can be improved with the “intelligent” selection of multiple PLS models. The quality of predictions for the respective external sets showed that the consensus models (CM) are better than the individual models (IM) in most of the cases. From the insights of the developed models, we concluded that attributes like a packed structure toward higher conductivity of electrons, auxiliary donor fragment of aromatic tertiary amines, number of thiophenes inducing the bathochromic shift and augmenting the absorption, presence of additional electron donors, enhancement of electron-donating abilities, number of non-aromatic conjugated C(sp<sup>2</sup>) which helps as conjugation extension units to broaden the absorption and highly conjugated  $\pi$ -systems exert positive contributions to the PCE. On the contrary, features negatively contributing to PCE are the followings: fragments which lower the tendency of localized  $\pi$ - $\pi^*$  transition, fragments related to larger volume and surface area of dyes along with hydrophobicity resulting in poor adhesion, fragment RC = N causing dye hydrolysis, steric hindrance for  $\pi$  electronic mobility, fragments enhancing polarity, etc. The identified features from the best QSPR model of the coumarin dataset was employed in designing of ten more efficient coumarin dyes (predicted %PCE ranging from 8.93 to 10.62) than the existing ones.

## 1. Introduction

A dye-sensitized solar cell (DSSC) is a molecular photovoltaic (PV) system that mimics nature's photosynthesis principle employing a dye to absorb solar radiant energy to generate charge carriers which are then separated, transported and collected as harnessed solar electricity [1]. In the last decade, the DSSCs have attracted considerable attention as an

alternative renewable energy source, and there is an extensive ongoing effort towards the design of organic dyes for DSSCs with high power conversion efficiency (PCE) to surpass some of the disadvantages of the previous inorganic solar cell systems like limiting weight, reducing cost, improving resources, and performing in an environment friendly manner [2]. In DSSCs, the dye acts as a photoactive component which converts photon to electricity through a series of stages where dye is

\* Corresponding author.

\*\* Corresponding author.

E-mail addresses: [kunalroy\\_in@yahoo.com](mailto:kunalroy_in@yahoo.com), [kunal.roy@jadavpuruniversity.in](mailto:kunal.roy@jadavpuruniversity.in) (K. Roy), [jerzy@icnanotox.org](mailto:jerzy@icnanotox.org) (J. Leszczynski).

<sup>1</sup> Both contributed equally as second authors.

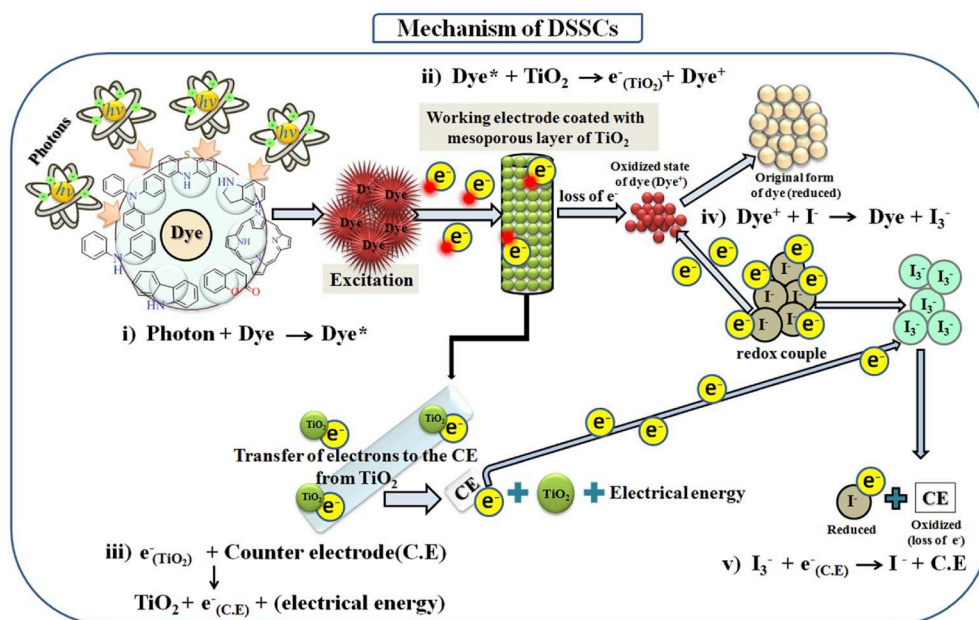


Fig. 1. General mechanism of Dye sensitized solar cells (DSSCs).

coated to the wide band gap semiconductor (in most of the cases TiO<sub>2</sub>) known as working electrode (Fig. 1). The DSSC is prepared with a mixture of high-performance electrolytes like iodide/triiodide (I<sup>-</sup>/I<sub>3</sub><sup>-</sup>) redox shuttle and additives [3]. The sensitizer is the key element as it controls the photon harvesting and electron transport into the nano-structured semiconductor surface and can offer a mean to govern the interfacial behavior of electron transfer at the TiO<sub>2</sub>/dye/electrolyte interface [4]. Most of the organic dyes in DSSCs are prepared based on the donor-acceptor (D-A)-like structure linked through a conjugated  $\pi$  spacer such as polyene and oligothiophene (D- $\pi$ -A) and usually have a rod like configuration. The electron donor units are composed of moieties like indoline, triarylamine, coumarin, etc., while carboxylic acid, cyanoacrylic acid, and rhodamine units are used as electron acceptors to fulfill the requirement [4]. The performance of DSSCs depends on important quantum properties like the highest occupied molecular orbital (HOMO), the lowest unoccupied molecular orbital (LUMO), and their distributions in the photosensitizers. Thus, dye's molecular structure and its orientation in the form of well-established high-performance D- $\pi$ -A structure framework [5] and/or D-A- $\pi$ -A structure [6] (where D and A stand for donor and acceptor fragments and  $\pi$  is the conjugated linker between D and A), which combines an auxiliary acceptor, improves their intramolecular charge transfers (ICT) and diminishes the optical band gap, are very important areas to study. Therefore, keeping constant all other components of DSSCs and just by changing the chemical structure of dyes, one can develop highly power conversion efficient solar system.

Dye-sensitizers can majorly be of two types: inorganic metal-based dye sensitizers and organic-metal free dye-sensitizers. The organic dyes are more environment friendly as well as easy to be modified structurally. Thus, a good number of organic dyes are explored for DSSCs such as triphenylamines [7,8], indolines [6], diketopyrrolopyrroles [9], anthocyanin dyes [10], perylenes [11], carbazole dyes [12], etc. In the current situation, ruthenium [13] and Zn-based DSSCs achieve PCE of 13% [14] experimentally, while organic-metal free dyes (triazatruxene (TAT) based D- $\pi$ -A dye) achieve PCE of 13.6% (as reported by Zhang et al. [15], which is the experimentally reported highest value at the present time). Interestingly, according to National Renewable Energy Laboratory solar cell efficiency chart, the DSSCs are the least performers among all existing solar cell types which shows the importance and requirement of more efficient and practical outcome-oriented

research in this specific field. As optimum performance of DSSCs depends on multi-layered aspects and components, thus a proper scheme needs to be followed for the design of dyes where all possible parameters of an efficient dye should be checked initially even before its synthesis. Over the last few years, low cost computational method like quantitative structure-property relationship (QSPR) is a recognized *in silico* tool in designing of potential dyes for DSSCs [16]. Instead of designing sensitizer dyes blindly spending a considerable amount of time and money, QSPR models can be a fruitful and reasonable approach to explore multiple chemical classes in search of the best possible dyes for DSSCs with higher PCE values than the existing ones in the market [17].

Multiple QSPR models have previously been explored for the designing of solar cell systems [18–25]. Fullerene-based polymer solar cell systems have been designed by our group for the very first time [18, 19] employing QSPR models and virtual screening strategically. Venkatraman et al. [20,21], Li et al. [22] and our group [23–25] are actively working in the designing of organic dyes for DSSCs from the least explored chemical classes like phenothiazines, indolines, tetrahydroquinolines, N,N'-dialkylanilines etc. We have already designed and proposed potentially efficient 'lead dyes' theoretically, employing QSPR models employing for 273 dyes followed by electrochemical and optoelectronic parameter evaluation which reported better (predicted) PCE values than the existing dyes, i.e. 18.88, 19.24, and 13.87 for tetrahydroquinolines, N,N'-dialkylanilines and indolines, respectively [23–25], inspiring synthesis and experimental studies of two efficient lead dyes from the N,N'-dialkylanilines family in future studies [17]. Venkatraman et al. [20] developed QSPR models based on eigenvalue (EVA) descriptors generated from vibrational frequencies which suggested the best goodness-of-fit as well as prediction capability. Venkatraman et al. [21] proposed effective *de novo* design employing QSPR analysis and projected five lead phenothiazine dyes where all designed compounds had predicted PCE values ranging from 9.2 to 9.52. Li et al. [22] developed cascaded QSPR models for the overall PCE using quantum chemical descriptors which successfully predicted the PCE for 354 organic dyes, offering a valuable tool for the design of future dye sensitizers with efficient PCE. It is obvious that the quantitative models not only predict the PCE of newly designed dyes, but also explore the structural as well as physicochemical features responsible for changes in PCE values which can be employed for future designing as well as alteration of structure scaffolds by experimentalists.

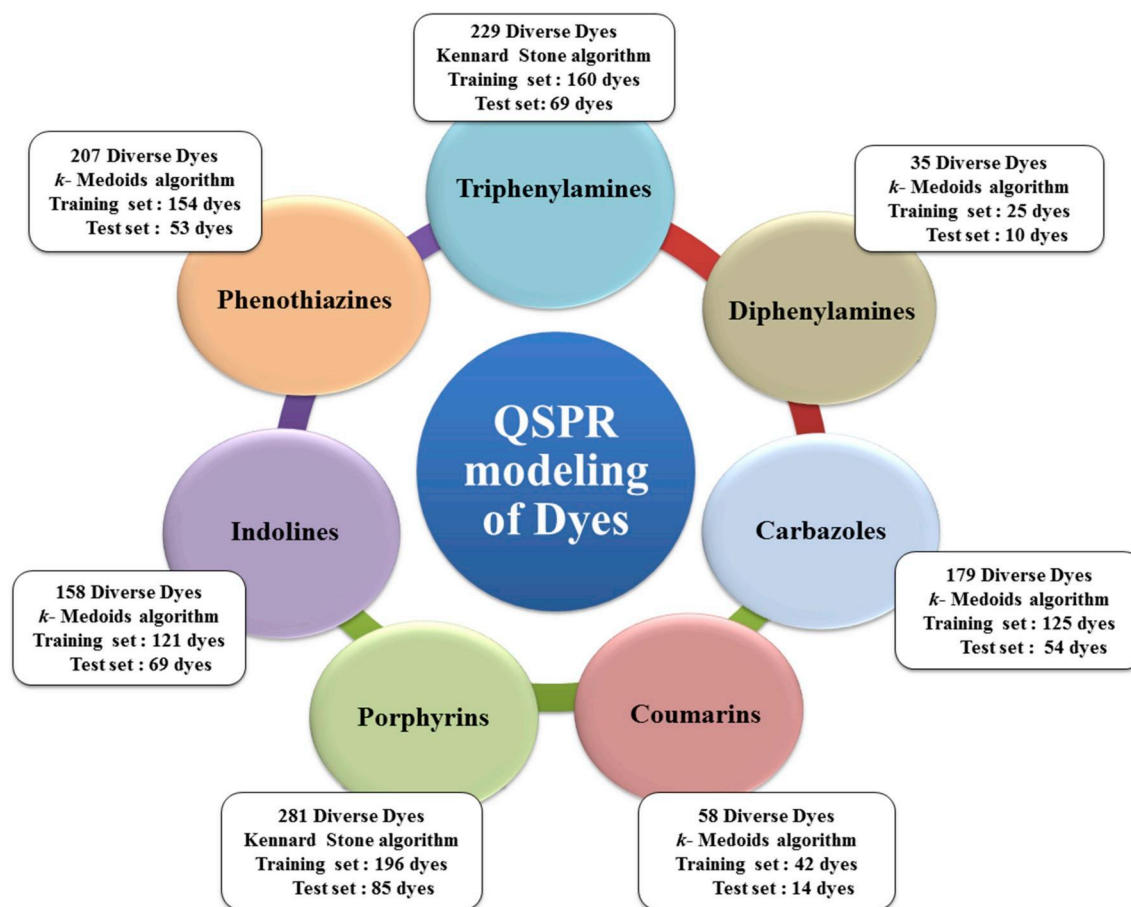


Fig. 2. Representation of various dye data sets used in QSPR modeling of dyes.

In this background, we have developed here multiple QSPR models for one of the biggest dye datasets for DSSCs consisting of around 1200 dyes covering seven important chemical classes, i.e. triphenylamines, phenothiazines, indolines, porphyrins, coumarins, carbazoles, and diphenylamines following all five principles of Organisation for Economic Co-operation and Development (OECD) for QSPR model development. The developed models should be significant tools for the prediction and screening of new and untested dye datasets as well as rich resources for the designing criteria of individual chemical classes, as each model has explored necessary structural fragments and molecular prerequisites for efficient dyes for DSSCs.

## 2. Materials and methods

### 2.1. Datasets

The experimental PCE values used for modeling of various chemical classes of dyes are obtained from the Dye Sensitized Solar Cell Database (DSSCDB) (<https://www.dyedb.com/>) [26]. Currently, the database is holding over 4000 experimental results for a diverse set of chemical classes. In a preliminary analysis of data, we have screened the dyes based on solar simulator (AM 1.5G 100 mW/cm<sup>2</sup>) and TiO<sub>2</sub> electrode. After that, we have divided the database into individual chemical classes of dyes; in this process, the mixture of dyes were discarded, the stand-alone chemical classes of dyes were separated for QSPR modeling. Since the response variable (PCE) refers to the energy terms, the modeling was performed without logarithmic conversion of the response (as customary in biological QSPR). Finally, around 1200 dyes were considered and classified into seven chemical classes and their experimental power conversion efficiency data were employed for QSPR

modeling. The seven datasets consist of 244 Triphenylamines (%PCE range: 0.053–10.1), 215 Phenothiazines (%PCE range: 0.12–8.18), 170 Indolines (%PCE range: 0.046–9.2), 300 Porphyrins (%PCE range: 0.0013–12.5), 56 Coumarins (%PCE range: 0.33–7.4), 179 Carbazoles (%PCE range: 0.038–12.5), and 35 diphenylamine (%PCE range: 0.4–8) dyes. The details of the data sets are provided in a [Supplementary Information](#) excel file.

### 2.2. Descriptor calculation

“The molecular descriptor is the ultimate result of a logical and mathematical operation which transfigures the chemical information encrypted within a symbolic representation of a molecule into the result of some regularized (standardized) experiments” [27]. The dye structures were drawn by using the Marvin Sketch 5.10.0 software [28]. Dragon software version 7 [29] and PaDEL-descriptor 2.21 software [30] were employed for the computation of 2D descriptors covering constitutional, ring descriptors, connectivity index, functional group counts, atom centered fragments, atom type E-states, 2D atom pairs, molecular properties (Dragon Software) and extended topochemical atom (ETA) indices descriptors (PaDEL-Descriptor software). To identify and interpret the structural fragments and physicochemical properties with ease and to avoid conformational complexity, we have employed only 2D descriptors.

### 2.3. Data set division

Individual datasets were divided into a training and a test set using Kennard-Stone (Triphenylamine, Porphyrin datasets) and “Modified k-medoid” [31] (Phenothiazine, Indoline, Coumarin, Carbazole, and



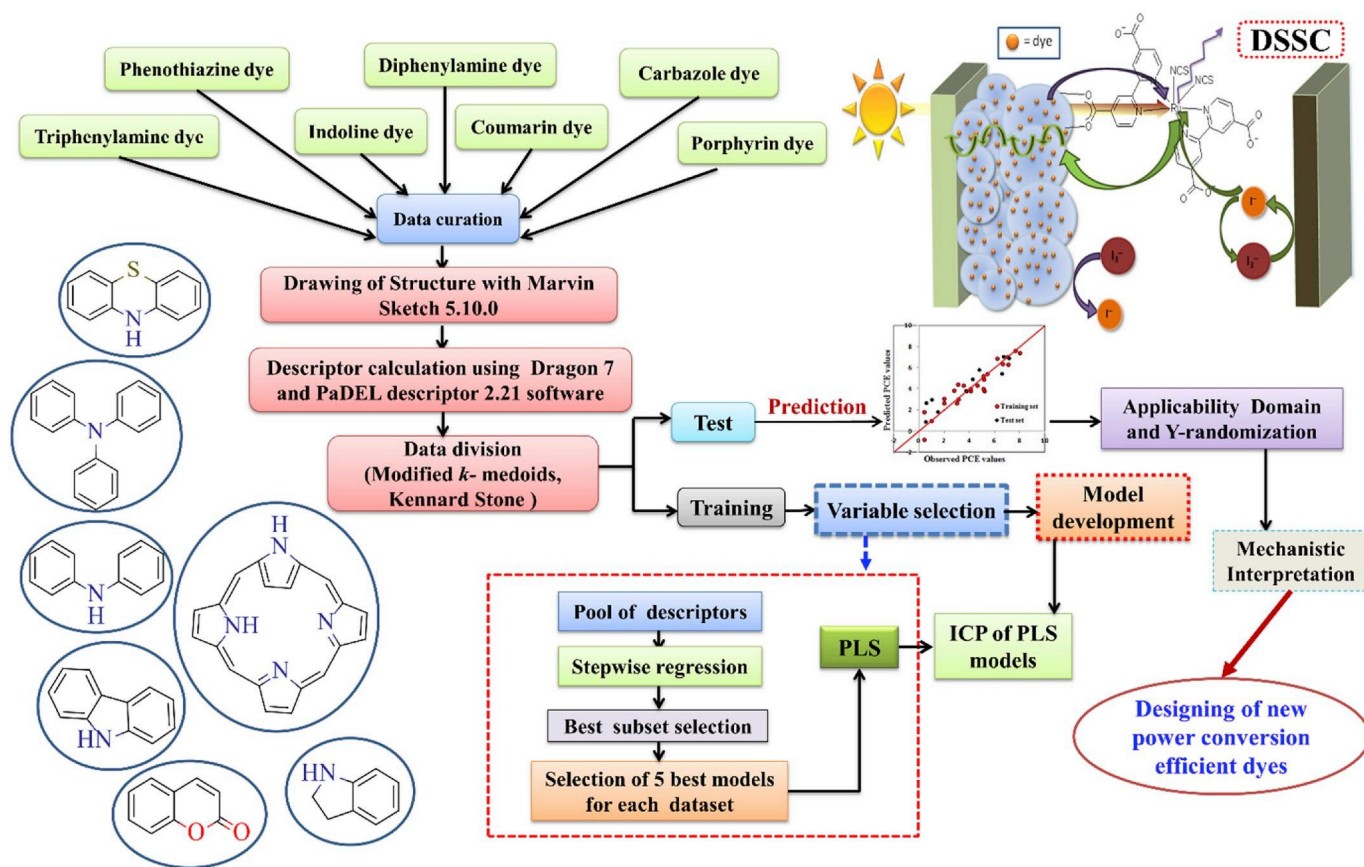


Fig. 3. Schematic representation of the steps involved in the development of QSPR models.

Diphenylamine datasets) algorithms. We have selected around 30% of dyes from each data set for the test set and the remaining 70% of compounds for the training set. The training set was used to develop the models, and the respective test set was used to validate the models for prediction purposes. The composition of training and test sets for the individual dataset is illustrated in Fig. 2.

#### 2.4. Variable selection and model development

The descriptor matrices of the training sets were pre-treated to eliminate intercorrelated descriptors from all the data sets followed by “multistage” stepwise regression analysis was performed to remove less important descriptors from the initial large pool of descriptors [32–36]. In this way, we have selected some manageable number of descriptors and made a reduced pool of descriptors for further processing. After that, we have run best features selection separately for all datasets using the reduced pool of descriptors employing ‘Best Subset selection v2.1 software’ [37]. We have developed multiple 13 descriptor models for Triphenylamines, 14 descriptor models for Phenothiazines, 13 descriptors models for Indolines, 12 descriptor models for Porphyrins, 5 descriptors models for Coumarins, 11 descriptor models for Carbazoles and 4 descriptor models for Diphenylamines datasets. Among all models obtained from the best subset selection, we have selected the best five models based on Mean Absolute Error (MAE) [38] values in case of each dataset. Descriptors selected in all five multiple linear regression (MLR) models for individual datasets were further subjected to partial least squares (PLS) modeling separately to refine the models in terms of predictivity and robustness. The MLR and PLS models were developed using MINITAB [39] and SIMCA-P software [40], respectively.

To explore whether the quality of predictions of external (test) compounds could be enhanced through an “intelligent” selection of multiple models (here, five models), we have further used our in house

Intelligent consensus predictor (ICP) tool [31,41]. The ICP software tool helps for finding and understanding the performance of consensus predictions in comparison to the quality obtained from the individual PLS models based on the MAE values (95%), as any single model may not be good for predictions for all test set compounds. The use of ICP is quite a rational approach for the prediction of test compounds considering multiple QSPR models developed from the same the training set. The steps involved in the development of PLS models are represented schematically in Fig. 3.

#### 2.5. Statistical validation metrics, AD study and Y-randomization test

For judging goodness-of-fit and predictive ability of the developed QSPR models, we have checked the statistical quality employing both internal and external validation metrics. Statistical parameters like determination coefficient ( $R^2$ ), explained variance ( $R_a^2$ ), leave-one-out cross-validated correlation coefficient ( $Q^2_{LOO}$ ), variance ratio (F), and standard error of estimate(s) were used to check the quality of training set fitting [42]. For external or test set validation,  $R^2_{pred}$  or  $Q^2F_1$  and  $Q^2F_2$  parameters were implemented [43]. We have also employed stringent validation metrics like  $r_m^2$  [43] and the mean absolute error (MAE) [38] values for both internal and external validation. The error based metrics were used to determine the true indication of the prediction quality in terms of prediction errors since they do not evaluate the performance of the model in comparison with the mean response [38]. Model Y-randomization test was performed using SIMCA-P software [40] to check whether the models are obtained by any chance or not. The applicability domain (AD) study was performed for each model using the DModX (distance to model X) approach at 99% confidence level using SIMCA-P software [40].

**Table 1**  
Statistical quality and validation parameters obtained from the developed PLS models.

Dataset	Type of model	Training set				Test set					AD criteria	Dixon Q-test	Euclidian distance
		R <sup>2</sup>	Q <sup>2</sup> <sub>(LOO)</sub>	$\overline{r}_{m(LOO)}^2$	$\Delta r_{m(LOO)}^2$	R <sup>2</sup> <sub>pred</sub> or Q <sup>2</sup> F <sub>1</sub>	Q <sup>2</sup> F <sub>2</sub>	$\overline{r}_{m(test)}^2$	$\Delta r_{m(test)}^2$	MAE <sub>(95%)</sub>			
Tri-phenyl amine	IM1 (LV:7)	0.67	0.60	0.47	0.19	0.60	0.60	0.47	0.20	1.06	NO	NO	0.4
	IM2 (LV:6)	0.68	0.61	0.48	0.20	0.59	0.59	0.46	0.24	1.03			
	IM3 (LV:5)	0.67	0.61	0.48	0.20	0.59	0.59	0.47	0.18	1.04			
	IM4 (LV:6)	0.65	0.59	0.46	0.22	0.58	0.58	0.44	0.24	1.05			
	IM5 (LV:7)	0.67	0.60	0.47	0.20	0.60	0.60	0.47	0.19	1.06			
	CM1	–	–	–	–	0.61	0.61	0.49	0.25	1.04			
Phenothiazine	CM2	–	–	–	–	0.62	0.61	0.49	0.20	1.05	NO	NO	0.4
	CM3	–	–	–	–	0.61	0.61	0.48	0.23	1.01			
	IM1 (LV:6)	0.70	0.64	0.51	0.20	0.69	0.69	0.58	0.18	0.91			
	IM2 (LV:6)	0.70	0.64	0.51	0.20	0.69	0.69	0.58	0.18	0.91			
	IM3 (LV:6)	0.70	0.63	0.51	0.20	0.70	0.70	0.60	0.13	0.85			
	IM4 (LV:6)	0.70	0.63	0.51	0.20	0.70	0.70	0.60	0.13	0.85			
	IM5 (LV:6)	0.70	0.63	0.51	0.20	0.70	0.70	0.60	0.15	0.86			
	CM0	–	–	–	–	0.70	0.70	0.60	0.16	0.88			
	CM1	–	–	–	–	0.71	0.71	0.60	0.18	0.87			
	CM2	–	–	–	–	0.71	0.71	0.61	0.18	0.87			
Indoline	CM3	–	–	–	–	0.73	0.73	0.63	0.18	0.83	NO	NO	0.4
	IM1 (LV:7)	0.74	0.66	0.55	0.16	0.72	0.72	0.67	0.12	0.72			
	IM2 (LV:7)	0.74	0.66	0.54	0.17	0.68	0.68	0.63	0.14	0.74			
	IM3 (LV:6)	0.73	0.66	0.55	0.17	0.67	0.67	0.64	0.18	0.68			
	IM4 (LV:7)	0.73	0.65	0.53	0.17	0.71	0.71	0.66	0.12	0.73			
	IM5 (LV:7)	0.75	0.67	0.56	0.16	0.68	0.68	0.64	0.13	0.76			
	CM0	–	–	–	–	0.71	0.71	0.66	0.13	0.71			
	CM1	–	–	–	–	0.71	0.71	0.66	0.12	0.70			
	CM2	–	–	–	–	0.71	0.71	0.66	0.13	0.69			
	CM3	–	–	–	–	0.74	0.74	0.69	0.15	0.66			
Porphyrin	IM1 (LV:6)	0.70	0.66	0.54	0.19	0.66	0.65	0.57	0.16	0.97	YES	YES	NO
	IM2 (LV:6)	0.70	0.66	0.54	0.19	0.66	0.66	0.57	0.16	0.96			
	IM3 (LV:6)	0.70	0.65	0.54	0.19	0.66	0.66	0.58	0.13	0.97			
	IM4 (LV:6)	0.70	0.65	0.54	0.19	0.67	0.66	0.58	0.13	0.96			
	IM5 (LV:5)	0.68	0.64	0.52	0.20	0.67	0.67	0.56	0.20	1.01			
	CM0	–	–	–	–	0.68	0.67	0.58	0.17	0.96			
	CM1	–	–	–	–	0.68	0.68	0.59	0.16	0.94			
	CM2	–	–	–	–	0.68	0.68	0.59	0.16	0.94			
	CM3	–	–	–	–	0.69	0.69	0.60	0.15	0.93			
	IM1 (LV:2)	0.78	0.71	0.61	0.15	0.60	0.58	0.43	0.29	0.95	0.4	NO	NO
Coumarin	IM2 (LV:2)	0.74	0.67	0.56	0.17	0.63	0.61	0.41	0.48	0.85			
	IM3 (LV:3)	0.75	0.67	0.57	0.15	0.61	0.59	0.36	0.33	0.89			
	IM4 (LV:2)	0.75	0.67	0.56	0.16	0.62	0.60	0.36	0.34	0.88			
	IM5 (LV:3)	0.71	0.65	0.54	0.19	0.68	0.66	0.53	0.24	0.84			
	CM0	–	–	–	–	0.65	0.63	0.42	0.30	0.88			
	CM1	–	–	–	–	0.63	0.61	0.40	0.31	0.92			
	CM2	–	–	–	–	0.63	0.61	0.40	0.31	0.92			
	CM3	–	–	–	–	0.61	0.63	0.37	0.34	0.89			
	IM1 (LV:5)	0.75	0.71	0.98	0.19	0.75	0.74	0.61	0.20	0.64	NO	NO	0.4
Carbazole	IM2 (LV:4)	0.75	0.70	0.99	0.20	0.73	0.71	0.58	0.21	0.65			
	IM3 (LV:5)	0.75	0.71	0.99	0.18	0.73	0.71	0.58	0.21	0.66			
	IM4 (LV:4)	0.74	0.70	1.01	0.19	0.74	0.72	0.57	0.21	0.64			
	IM5 (LV:4)	0.74	0.69	1.01	0.20	0.73	0.71	0.55	0.22	0.67			
	CM0	–	–	–	–	0.75	0.73	0.58	0.21	0.63			
	CM1	–	–	–	–	0.75	0.73	0.58	0.20	0.63			
	CM2	–	–	–	–	0.75	0.73	0.58	0.20	0.63			
	CM3	–	–	–	–	0.75	0.73	0.58	0.21	0.63			
	IM1 (LV:3)	0.88	0.81	0.74	0.01	0.83	0.83	0.62	0.15	0.65	NO	NO	NO
Di-phenyl amine	IM2 (LV:2)	0.86	0.81	0.74	0.03	0.74	0.73	0.49	0.24	0.73			
	IM3 (LV:2)	0.87	0.82	0.74	0.12	0.83	0.82	0.70	0.14	0.77			
	IM4 (LV:2)	0.87	0.81	0.74	0.09	0.81	0.80	0.65	0.16	0.92			
	IM5 (LV:2)	0.88	0.82	0.76	0.04	0.80	0.79	0.57	0.19	0.65			
	CM0	–	–	–	–	0.84	0.83	0.62	0.16	0.61			
	CM1	–	–	–	–	0.84	0.83	0.62	0.16	0.61			
	CM2	–	–	–	–	0.84	0.83	0.63	0.15	0.61			
	CM3	–	–	–	–	0.85	0.84	0.74	0.12	0.65			

LV: Latent variable for PLS models; CM0: Ordinary consensus predictions; CM1: Average of predictions from ‘qualified’ Individual models; CM2: Weighted average predictions from ‘qualified’ Individual models; CM3: Best selection of predictions compound wise from ‘qualified Individual models’; Best model for individual dataset is marked in bold.

### 3. Results and discussion

Statistically acceptable and robust individual (IM), as well as consensus models (CM), were developed as depicted in Table 1. Analyzing the obtained results, we found that in most of the cases,

consensus predictions of multiple PLS models were better than the results obtained from the individual PLS models. From among all validation metrics, we have selected the best models based on MAE<sub>(95%)</sub> to give more importance on the prediction error of test or external compounds. The CM3 model which signifies ‘the best selection of compound wise

**Box 1**

IM1

$$\text{PCE} = 2.971 - 48.538 \times \text{GD} - 0.599 \times \text{F06}[\text{N}-\text{O}] + 1.587 \times \text{B09}[\text{C}-\text{S}] - 0.298 \times \text{SdssC} + 3.545 \times \text{B06}[\text{C}-\text{O}] - 2.415 \times \text{nN}(\text{CO})_2 - 2.992 \times \text{NdsN} - 2.805 \times (\text{C}-038) + 4.578 \times (\text{nRC}=\text{N}) - 0.685 \times \text{F05}[\text{N}-\text{N}] - 2.257 \times \text{B07}[\text{O}-\text{S}] - 2.597 \times (\text{C}-043) + 1.465 \times \text{B06}[\text{O}-\text{S}]$$

IM2

$$\text{PCE} = -1.208 - 56.313 \times \text{GD} - 0.310 \times \text{SdssC} - 0.703 \times \text{F06}[\text{N}-\text{O}] + 2.597 \times \text{B06}[\text{C}-\text{O}] + 5.304 \times (\text{nRC}=\text{N}) - 2.849 \times \text{NdsN} - 3.098 \times (\text{C}-038) - 2.201 \times \text{nN}(\text{CO})_2 + 85.425 \times \text{X4Av} - 0.654 \times \text{F05}[\text{N}-\text{N}] - 2.032 \times \text{B07}[\text{O}-\text{S}] - 2.817 \times (\text{C}-043) + 11.530 \times \text{ETA\_Shape\_Y}$$

IM3

$$\text{PCE} = 1.495 - 55.860 \times \text{GD} - 0.329 \times \text{SdssC} - 0.622 \times \text{F06}[\text{N}-\text{O}] + 3.116 \times \text{B06}[\text{C}-\text{O}] - 2.996 \times \text{NdsN} + 2.403 \times \text{nN}(\text{CO})_2 + 88.416 \times \text{X4Av} - 3.061 \times (\text{C}-038) + 4.744 \times (\text{nRC}=\text{N}) - 2.487 \times \text{B07}[\text{O}-\text{S}] - 0.556 \times \text{F05}[\text{N}-\text{N}] - 2.786 \times (\text{C}-043) + 1.632 \times \text{B06}[\text{O}-\text{S}]$$

IM4

$$\text{PCE} = 1.424 - 54.789 \times \text{GD} - 0.749 \times \text{B02}[\text{N}-\text{O}] - 0.511 \times \text{F06}[\text{N}-\text{O}] - 1.574 \times \text{nN}(\text{CO})_2 - 0.271 \times \text{SdssC} + 3.385 \times \text{B06}[\text{C}-\text{O}] + 83.242 \times \text{X4Av} - 3.152 \times \text{NdsN} + 4.815 \times (\text{nRC}=\text{N}) - 1.894 \times \text{B07}[\text{O}-\text{S}] - 2.657 \times (\text{C}-038) - 0.595 \times \text{F05}[\text{N}-\text{N}] - 2.570 \times (\text{C}-043)$$

IM5

$$\text{PCE} = 2.971 - 48.538 \times \text{GD} - 0.599 \times \text{F06}[\text{N}-\text{O}] + 1.587 \times \text{B09}[\text{C}-\text{S}] - 0.298 \times \text{SdssC} + 3.545 \times \text{B06}[\text{C}-\text{O}] - 2.415 \times \text{nN}(\text{CO})_2 - 2.992 \times \text{NdsN} - 2.805 \times (\text{C}-038) + 4.578 \times (\text{nRC}=\text{N}) - 0.685 \times \text{F05}[\text{N}-\text{N}] - 2.257 \times \text{B07}[\text{O}-\text{S}] - 2.597 \times (\text{C}-043) + 1.465 \times \text{B06}[\text{O}-\text{S}]$$

predictions from the selected individual models' is the winner model for following datasets: tri-phenylamines, phenothiazines, indolines, and porphyrins. However, all four consensus models evolved as the winner model in case of the carabazole dataset, whereas CM0 (ordinary consensus predictions), CM1 (average of predictions from 'qualified' Individual models) and CM2 (weighted average predictions from 'qualified' Individual models) models are winner for the diphenylamine dataset. In contrast, in case of the coumarin dataset, individual model 5 (IM5) is the best model. It is quite evident from the outcome that predictability of consensus models is much better than the individual models; the former not only nullifies the error of predictions from an individual model but also enhances the reliability of the predictions for the true external dataset. All the individual models are mechanistically interpreted based on the modeled descriptors. In case of all the datasets, we have selected five PLS models in each case based on MAE values followed by development of consensus models. The selected models contain 13, 14, 13, 12, 5, 11 and 4 descriptors for the triphenylamine, phenothiazine, indoline, porphyrin, coumarin, carbazole, and diphenylamine datasets, respectively. To understand the order of significance of the modeled descriptors or variables in a descending order, we have prepared Variable Importance in Projection (VIP) plot for individual models of each dataset which can be found in Figs. S1–S7 in the Supplementary materials file.

### 3.1. Dataset 1: modeling of PCE property of triphenylamine dyes

The significant descriptors obtained from the five PLS models are indicated in Box 1. The mechanistic interpretation of all the descriptors is discussed below.

- i) The atom type E-state descriptor NdsN represents number of nitrogen atoms with double and single bonds (=N-) contributing negatively to the PCE indicating that presence of this feature in the dye may decrease the PCE as reflected for following examples: **40** (NdsN = 1; PCE = 0.44), **161** (NdsN = 1; PCE = 0.45) and **144** (NdsN = 1; PCE = 0.18) and *vice versa* in case of dyes like **97** (NdsN = 0; PCE = 7.83), **204** (NdsN = 0; PCE = 8.06) and **238** (NdsN = 0; PCE = 10.1). This fragment lowers the tendency of localized  $\pi$ - $\pi^*$  transition due to intramolecular charge transfer transition (ICT) from the triphenylamine donor. As a result, the

absorption maxima will decrease and thereby the PCE values may decrease [44].

- ii) The 2D atom pair descriptor B06[C–O] denotes presence/absence of carbon and oxygen atoms at the topological distance 6, which contributed positively towards the PCE due to its positive regression coefficients. Thus, presence of this fragment in the dye molecules may increase the PCE property as shown in dyes **162** (B06[C–O] = 1; PCE = 7.78), **178** (B06[C–O] = 1; PCE = 6.95) and **189** (B06[C–O] = 1; PCE = 6.82) and *vice versa* in case of dyes **166** (B06[C–O] = 0; PCE = 0.087), **169** (B06[C–O] = 0; PCE = 0.058) and **171** (B06[C–O] = 0; PCE = 0.093). Presence of this group in the dye molecules leads to bathochromic shift of the absorption spectrum and enhancement of the molar extinction coefficient of the dye which directly causes higher PCE [45].
- iii) The 2D atom pair descriptor B07[O–S] describes the presence/absence of oxygen and sulfur atoms at the topological distance 7, contributing negatively towards the PCE. Due to the presence of this fragment in the donor groups, there is a narrowing of the absorption range of dyes which causes latency decrease of rapid  $\pi$ -conjugation [46]. Eventually, the PCE values decrease as shown in the dyes **63** (B07[O–S] = 1, PCE = 0.77), **67** (B07[O–S] = 1; PCE = 0.45) and **144** (B07[O–S] = 1; PCE = 0.18). In contrast, the dyes having no such fragments may experience an enhancement in the PCE property as shown in dyes **14** (B07[O–S] = 0, PCE = 6.22), **95** (B07[O–S] = 0, PCE = 6.51) and **107** (B07[O–S] = 0, PCE = 7.67).
- iv) Another 2D atom pair descriptor B09[C–S] denotes the presence/absence of carbon and sulfur atoms at the topological distance 9 which signifies longer chain in a molecule where sulfur atom is a part of thiophene ring resulting in a slight hypsochromic and a hypochromic effect in the ICT band. This may be explained by the steric hindrance induced by the branched-chain that increases the torsion angle between the triphenylamine moiety and the thiophene unit. This torsion impedes good delocalization of the  $\pi$  electrons and blue-shifts the position of the ICT band and augment the absorption [47]. Thus, the dyes having such fragment may have an enhanced PCE property as shown in dyes **14** (B09[C–S] = 1, PCE = 6.22), **94** (B09[C–S] = 1, PCE = 7.03) and **95** (B09[C–S] = 1, PCE = 6.51) and *vice versa* in case of dyes **8**



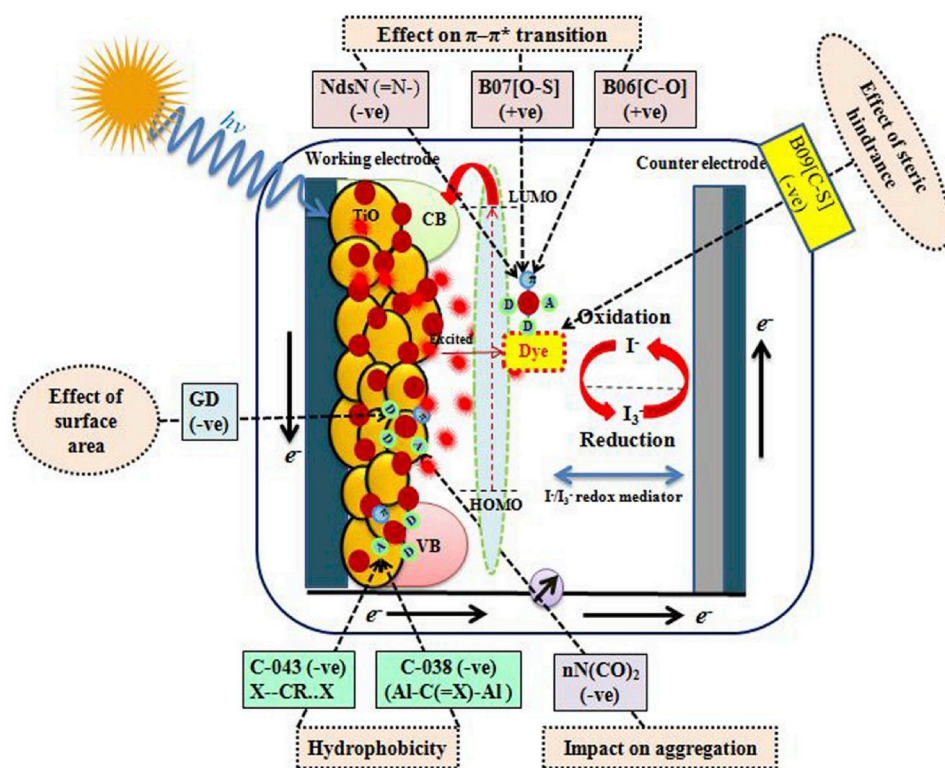


Fig. 4. Contribution of features on controlling PCE values of the Triphenylamine dyes.

- (B09[C-S] = 0, PCE = 0.08), **9** (B09[C-S] = 0, PCE = 0.83) and **40** (B09[C-S] = 0, PCE = 0.44).
- v) The atom-centered fragment descriptor C-038 represents the Al-C(=X)-Al fragment (where, Al: aliphatic groups and X: any electronegative atoms like O, N, S, P, Se, halogens) while another descriptor C-043 represents X-CR..X (R: any group linked through carbon) [48,49]. Both descriptors contributed negatively towards the PCE property by contributing to hydrophobicity and acting as a buffer between the semiconductor and the electrolyte, thus effectively preventing back-transfer of electrons from the semiconductor's conduction band to the redox couple. Therefore, the charge recombination is reduced [50]. Thus, the PCE values may decrease in the dyes containing these fragments as shown in dyes like **68** (C-038 = 2, PCE = 0.95), **69** (C-038 = 2; PCE = 1.1) and **217** (C-038 = 1; PCE = 2.03) for C-038 descriptor; and **49** (C-043 = 2; PCE = 2.98), **72** (C-043 = 1, PCE = 3.34) and **91** (C-043 = 1, PCE = 1.02) for C-043 descriptor. On the other hand, the dyes having no such fragments may have enhanced PCE property as shown in dyes **97** (C-038 = 0 & C-043 = 0; PCE = 7.83), **204** (C-038 = 0 & C-043 = 0; PCE = 8.06) and **238** (C-038 = 0 & C-043 = 0; PCE = 10.1).
- vi) The functional group count descriptor nN(CO) represents the number of imides(thio) in the dye structures. The negative regression coefficient of this descriptor indicates that presence of this fragment in the dye molecules may decrease the PCE property as shown in dyes **148** (nN(CO)<sub>2</sub> = 1; PCE = 0.18) and **170** (nN(CO)<sub>2</sub> = 2; PCE = 0.053) and *vice versa* in case of dyes **113** (nN(CO)<sub>2</sub> = 0; PCE = 7.21), **175** (nN(CO)<sub>2</sub> = 0; PCE = 7.28) and **234** (nN(CO)<sub>2</sub> = 0; PCE = 7.25). Presence of this feature favours dye hydrolysis which improves the aggregation property of the dye over the TiO<sub>2</sub> surface and improves the recombination reaction between redox electrolyte and electrons in the TiO<sub>2</sub> nanolayer. As a result, the linkage will be distorted and thereby the PCE values may decrease [51].

- vii) The ETA\_Shape\_Y descriptor deals with size and branching in the molecular structure. This descriptor contributes positively towards the PCE property as indicated by its positive regression coefficient. The higher numerical value of this descriptor may enhance the bulk of dyes resulting in sensitized wide-bandgap in the nanostructured photoelectrode [52]. The PCE values may increase with an increase of this descriptor value as shown in the dyes **48** (ETA\_Shape\_Y = 0.366; PCE = 6.01), **120** (ETA\_Shape\_Y = 0.342; PCE = 7.66) and **179** (ETA\_Shape\_Y = 0.360; PCE = 7.58). On the other hand, the lower numerical value of this descriptor may decrease the PCE property as shown in dyes **40** (ETA\_Shape\_Y = 0.235; PCE = 0.44), **115** (ETA\_Shape\_Y = 0.241; PCE = 0.6) and **134** (ETA\_Shape\_Y = 0.170; PCE = 1.7).
- viii) Graph density (GD) is derived from the H-depleted molecular graph and calculated from the following formula:

$$GD = \frac{2 \cdot nBo}{nSK \cdot (nSK - 1)}$$

Here, nBo is the number of graph edges (i.e., non-H bonds) and nSK is the number of vertices in the graph (i.e., non-H atoms). This descriptor indicates the surface area of the dye which leads to prolongation of the electron injection into the nano-structured TiO<sub>2</sub> [53]. Thus, higher surface area may decrease the PCE property of dyes in DSSC as evident from the negative contribution. The higher numerical values of this descriptor may decrease the PCE property as shown in dyes **8** (GD = 0.104; PCE = 0.08), **9** (GD = 0.098; PCE = 0.83) and **115** (GD = 0.133; PCE = 0.6) and *vice versa* in case of dyes **143** (GD = 0.019; PCE = 6.6), **204** (GD = 0.017; PCE = 8.06) and **243** (GD = 0.0189; PCE = 6.69).

- ix) The positive regression coefficients of 2D atom pair descriptor B06 [O-S] (presence or absence of oxygen and sulfur atoms at the topological distance 6) and the connectivity index X4Av (average valence connectivity index of order 4) as well as the functional group count descriptor nRC = N (number of aliphatic imines) indicate that presence of these fragments in the triphenylamine

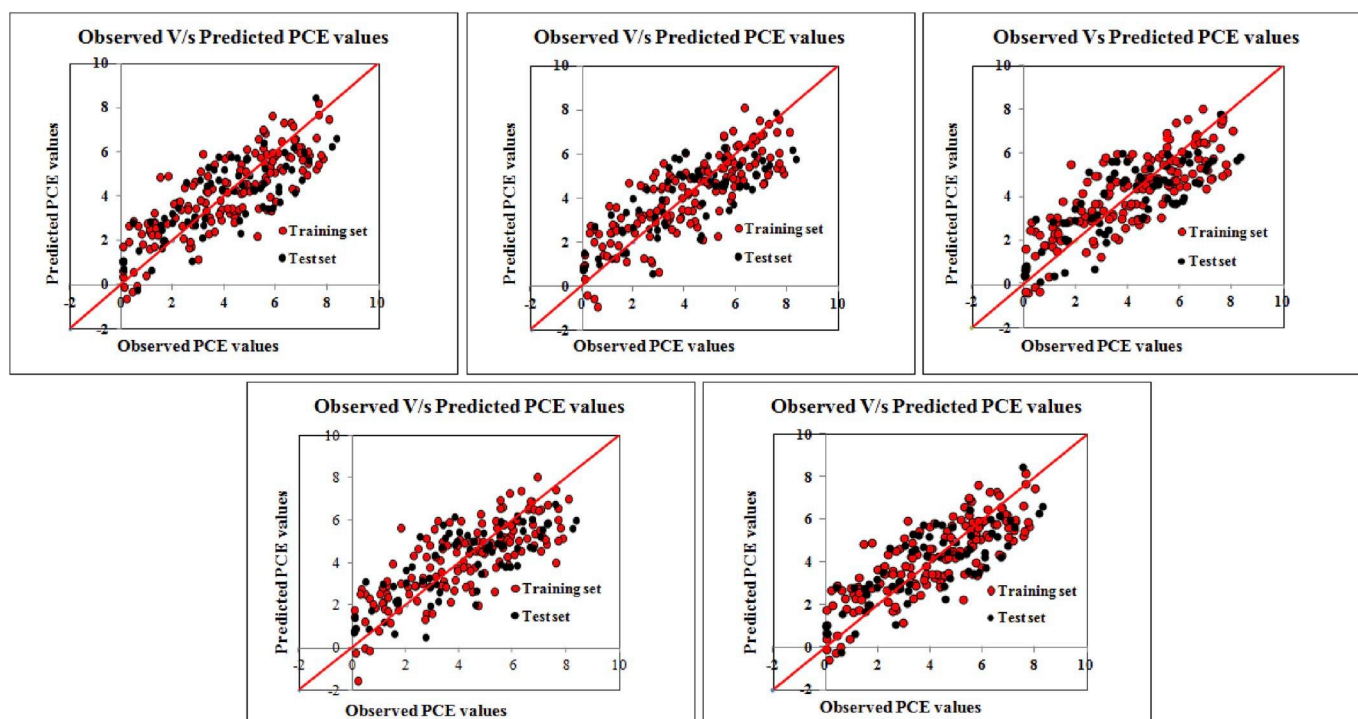


Fig. 5. Scatter plots of the observed and the predicted PCE values of the developed PLS models (IM1-IM5).

dye molecules may enhance the PCE property as shown in dyes **107** (B06[O-S] = 1; PCE = 7.67), **179** (B06[O-S] = 1; PCE = 7.58) and **122** (B06[O-S] = 1; PCE = 5.5) in case of B06[O-S] descriptor; **88** (X4Av = 0.072; PCE = 6.65), **89** (X4Av = 0.067; PCE = 7.6) and **239** (X4Av = 0.062; PCE = 6.91) for X4Av descriptor; and **70** (nRC = N = 2; PCE = 3.3) and **71** (nRC = N = 2; PCE = 3.9) for nRC = N descriptor and *vice versa* in case of dyes **7** (B06[O-S] = 0; PCE = 1.78), **39** (B06[O-S] = 0; PCE = 1.49) and **91** (B06[O-S] = 0; PCE = 1.02); **115** (X4Av = 0.028; PCE = 0.6), **119** (X4Av = 0.029; PCE = 1.1) and **159** (X4Av = 0.028; PCE = 0.3); **144** (nRC = N = 0; PCE = 0.18) and **166** (nRC = N = 0; PCE = 0.087), respectively.

x) The negative regression coefficient of 2D atom pair descriptors like F05[N-N] (frequency of two nitrogen atoms at the topological distance 5), F06[N-O] (frequency of nitrogen and oxygen atoms at topological distance 6), B02[N-O] (presence or absence of nitrogen and oxygen atoms at the topological distance 2) and the atom type E-state descriptor SdssC (sum of double, single and single bonded carbon E-states) (=C<) indicate that presence of these fragments in triphenylamine dye molecules may decrease the PCE property as shown in dyes **149** (F05[N-N] = 1; PCE = 2.6), **156** (F05[N-N] = 1; PCE = 2.2) and **159** (F05[N-N] = 1; PCE = 0.3) for F05[N-N]; **63** (PCE = 0.77), **166** (PCE = 0.087) and **171** (PCE = 0.093) for F06[N-O], **67** (PCE = 0.45), **87** (PCE

## Box 2

### IM1

$$\text{PCE} = 61.755 + 2.021 \times \text{B04}[\text{N}-\text{O}] - 2.604 \times \text{B04}[\text{O}-\text{S}] - 0.719 \times (\text{H}-052) + 1.727 \times \text{B08}[\text{C}-\text{O}] + 1.552 \times \text{B07}[\text{N}-\text{S}] + 0.438 \times \text{O}\% - 103.322 \times \text{Mi} + 59.547 \times \text{X0A} + 3.926 \times \text{totalcharge} - 1.520 \times \text{F09}[\text{S}-\text{S}] + 1.348 \times \text{B10}[\text{C}-\text{S}] - 0.127 \times \text{F05}[\text{C}-\text{O}] + 0.002 \times \text{D/Dtr05} + 0.370 \times (\text{C}-022)$$

### IM2

$$\text{PCE} = 61.755 + 2.021 \times \text{B04}[\text{N}-\text{O}] - 2.604 \times \text{B04}[\text{O}-\text{S}] - 0.719 \times (\text{H}-052) + 1.727 \times \text{B08}[\text{C}-\text{O}] + 1.552 \times \text{B07}[\text{N}-\text{S}] + 0.438 \times \text{O}\% - 103.322 \times \text{Mi} + 59.547 \times \text{X0A} + 3.926 \times \text{totalcharge} - 1.520 \times \text{F09}[\text{S}-\text{S}] + 1.348 \times \text{B10}[\text{C}-\text{S}] - 0.127 \times \text{F05}[\text{C}-\text{O}] + 0.002 \times \text{D/Dtr05} + 0.370 \times \text{nR}\#\text{C}-$$

### IM 3

$$\text{PCE} = 64.630 + 2.348 \times \text{B04}[\text{N}-\text{O}] - 2.711 \times \text{B04}[\text{O}-\text{S}] + 1.632 \times \text{B08}[\text{C}-\text{O}] + 1.561 \times \text{B07}[\text{N}-\text{S}] - 102.274 \times \text{Mi} + 0.418 \times \text{O}\% + 0.978 \times \text{F04}[\text{O}-\text{O}] + 3.789 \times \text{totalcharge} - 0.133 \times \text{F05}[\text{C}-\text{O}] + 53.834 \times \text{X0A} - 1.530 \times \text{F09}[\text{S}-\text{S}] + 0.002 \times \text{D/Dtr05} + 0.424 \times \text{nR}\#\text{C}- + 1.243 \times \text{B10}[\text{C}-\text{S}]$$

### IM4

$$\text{PCE} = 64.630 + 2.348 \times \text{B04}[\text{N}-\text{O}] - 2.711 \times \text{B04}[\text{O}-\text{S}] + 1.632 \times \text{B08}[\text{C}-\text{O}] + 1.561 \times \text{B07}[\text{N}-\text{S}] - 102.274 \times \text{Mi} + 0.418 \times \text{O}\% + 0.978 \times \text{F04}[\text{O}-\text{O}] + 3.789 \times \text{totalcharge} - 0.133 \times \text{F05}[\text{C}-\text{O}] + 53.834 \times \text{X0A} - 1.530 \times \text{F09}[\text{S}-\text{S}] + 0.002 \times \text{D/Dtr05} + 0.424 \times (\text{C}-022) + 1.243 \times \text{B10}[\text{C}-\text{S}]$$

### IM5

$$\text{PCE} = 63.969 + 2.333 \times \text{B04}[\text{N}-\text{O}] - 2.582 \times \text{B04}[\text{O}-\text{S}] + 1.568 \times \text{B08}[\text{C}-\text{O}] + 1.537 \times \text{B07}[\text{N}-\text{S}] - 104.701 \times \text{Mi} + 0.427 \times \text{O}\% + 4.242 \times \text{totalcharge} + 58.245 \times \text{X0A} - 1.897 \times \text{F09}[\text{S}-\text{S}] + 0.002 \times \text{D/Dtr05} + 0.424 \times (\text{C}-022) - 0.117 \times \text{F05}[\text{C}-\text{O}] + 1.202 \times \text{B10}[\text{C}-\text{S}] + 0.824 \times \text{B07}[\text{S}-\text{S}]$$



= 0.63) and **103** (PCE = 1.12) for B02[N-O] descriptor; and **66** (SdssC = 4.65, PCE = 0.45) and **67** (SdssC = 3.08; PCE = 0.27) for SdssC descriptor. On other hand, absence of these fragments may be important for the PCE property of dyes as shown in the dyes like **162** (PCE = 7.78), **172** (PCE = 7.02) and **175** (PCE = 7.28) for F05[N-N] descriptor; **107** (PCE = 7.67), **113** (PCE = 7.21) and **177** (PCE = 6.95) for F06[N-O] descriptor; **97** (PCE = 7.83), **189** (PCE = 6.82) and **204** (PCE = 8.06) for B02[N-O] descriptor, **195** (PCE = 6.78), **205** (PCE = 6.57) and **231** (PCE = 7.67) for SdssC descriptor.

The mechanistic interpretation of the triphenylamine dyes from all models is schematically portrayed in Fig. 4. The scatter plots of observed vs. predicted PCE property related to the triphenylamine dyes of DSSCs for all the PLS models are shown in Fig. 5.

### 3.2. Dataset 2: modeling of PCE property of phenothiazine dyes

The significant descriptors along with the mathematical equations of the five PLS models are illustrated in Box 2. The modeled descriptors are discussed below in detail with their meaning along with how they influence the PCE values.

The constitutional descriptor total charge is defined as sum of the charges of the individual atoms which contributes positively towards the PCE property as indicated by its positive regression coefficient. The dyes with a higher value of the descriptor may push forward the  $\pi$ - $\pi^*$  transitions leading to the efficient ICT in the donor groups of dyes resulting in an increase in the PCE values [54]. Thus, the dyes bearing higher charges atoms may have higher PCE property as shown in the dyes **191** (Total charge = 1; PCE = 7.1) and **192** (Total charge = 1; PCE = 6.9). On the other hand, the dyes having lower charges may have low PCE property as evidenced from the dyes **62** (Total charge = 0; PCE = 0.4), **66** (Total charge = 0; PCE = 1.8) and **70** (Total charge = 0; PCE = 0.6).

The 2D atom pair descriptor B07[N-S] indicates the presence or absence of the nitrogen and sulfur atoms at the topological distance 7. The positive regression coefficient of this descriptor indicates presence of these two atoms at the topological distance 7 in the dye may increase the PCE as evidenced by dyes **51** (B07[N-S] = 1; PCE = 7.7), **141** (B07[N-S] = 1; PCE = 7.87) and **180** (B07[N-S] = 1; PCE = 6.64) and *vice versa* in case of dyes **114** (B07[N-S] = 0; PCE = 1.12), **155** (B07[N-S] = 0; PCE = 0.93) and **157** (B07[N-S] = 0; PCE = 0.12). Presence of nitrogen and sulfur atoms at the topological distance 7 improves the photo-excitation by increasing the localized  $\pi$ - $\pi^*$  transition of the dye [46]. The photo excitation of the dye increases the PCE property.

The constitutional descriptor O% denotes the percentage of oxygen atoms and has a positive contribution to the PCE property as evident from the dyes **78** (O% = 11.764; PCE = 6.72), **197** (O% = 11.764; PCE = 6.09) and **199** (O% = 10; PCE = 6.58). On the other hand, the dyes containing low number of oxygen atoms may decrease the PCE property as shown in dyes **111** (O% = 0; PCE = 0.7) and **112** (O% = 0; PCE = 1.06). Oxygen atoms are involved in the conduction of electrons (towards the excitation state) and they have a natural tendency to form a closely packed structure (<https://neutronsources.org/news/scientific-highlights/neutron-power-finding-useful-oxygen-atoms-and-ions.html>) [55]. As a result, the higher conductivity carriers in the dyes enhance the PCE property.

The 2D atom pair descriptor B08[C-O] indicates the presence/absence of carbon and oxygen atoms at the topological distance 8 with positive effects to the PCE property as shown in the dyes **5** (B08[C-O] = 1; PCE = 6.98), **122** (B08[C-O] = 1; PCE = 6.82), **142** (B08[C-O] = 1; PCE = 7.98) and **158** (B08[C-O] = 1; PCE = 7.33) and *vice versa* in case of dyes **27** (B08[C-O] = 0; PCE = 1.88), **98** (B08[C-O] = 0; PCE = 2.66) and **112** (B08[C-O] = 0; PCE = 1.06). Presence of carbon and oxygen atoms at the topological distance 8 signifies the effect of donor and additional donor through a linkage in the dye system. This specific structural fragment helps to achieve absorption band broadening which

influences the PCE property of dye molecules [54]. Thus, the broadening of the absorption band increases the PCE property of dye molecules in the solar cell system.

Another 2D atom pair descriptor B04[N-O] is defined as the presence or absence of nitrogen and oxygen atoms at the topological distance 4 which positively contributes to the PCE property as evident from the dyes **12** (B04[N-O] = 1; PCE = 6.29), **127** (B04[N-O] = 1; PCE = 6.87) and **129** (B04[N-O] = 1; PCE = 8.08). On the contrary, dyes like **56** (B04[N-O] = 0; PCE = 0.73), **57** (B04[N-O] = 0; PCE = 0.33) and **70** (B04[N-O] = 0; PCE = 0.6) have low PCE values. This descriptor signifies the distance between the nitrogen and the oxygen atoms which is referred to a strong cyano acceptor and a chelating anchoring mode of the carboxylation which plays a crucial role to regulate the PCE property of dyes [56]. Thus, the presence of strong acceptor and chelating anchors in dye leads to higher PCE values.

The connectivity descriptor X0A denotes average connectivity index of order 0 and the ring descriptor, D/Dtr05, states distance/detour ring index of order 5. These descriptors have an impact on the surface area of the dye molecules. The positive regression coefficients of these descriptors indicated that the dyes having large surface areas may have higher PCE property as shown in the dyes **12** (X0A = 0.713; PCE = 6.29), **63** (X0A = 0.728; PCE = 6.8) and **126** (X0A = 0.708; PCE = 7.44) (in case of X0A descriptor) and other dyes like **10** (D/Dtr05 = 807; PCE = 6.67), **75** (D/Dtr05 = 674; PCE = 7.3) and **186** (D/Dtr05 = 782; PCE = 7.94) (in case of D/Dtr05 descriptor). On the contrary, lower numerical values of these descriptors may reduce the PCE value of dye molecules as shown in dyes like **27** (X0A = 0.696, D/Dtr05 = 0; PCE = 1.83), **56** (X0A = 0.701, D/Dtr05 = 0; PCE = 0.99) and **57** (X0A = 0.697, D/Dtr05 = 0; PCE = 0.73). Large surface area of the dyes may affect the photon capturing ability due to the sensitized wide band gap in the photo electrode [52]. As a result of the sensitized wide band gap in the nano-structured photo electrode, the PCE values may be enhanced.

The atom centered fragment descriptor H-052 denotes H<sup>+</sup> attached to C0(sp<sup>3</sup>) with 1X attached to next carbon (where, X: any electronegative atom O, N, S, P, Se, halogens; the superscript <sup>c</sup> represents the formal oxidation number) which has a negative contribution to the PCE. Thus, the dyes bearing this fragment may have lower PCE values as shown in the dyes **56** (H-052 = 2; PCE = 0.73), **57** (H-052 = 2; PCE = 0.99) and **65** (H-052 = 1; PCE = 1.3), whereas the dyes **141** (H-052 = 0; PCE = 7.87), **142** (H-052 = 0; PCE = 7.98) and **143** (H-052 = 0; PCE = 8.06) showed higher PCE values as these dyes are devoid of this fragment. The presence of this fragment favors the lipophilicity of the dyes which causes alterations in the energy cascade as a result of the physicochemical changes occurring in the dyes [50]. Thus, the physicochemical alterations such as poor solubility on semiconductors' porous layer decreases the PCE values.

The constitutional descriptor Mi represents the mean first ionization potential (scaled on a carbon atom) which shows a negative contribution to the PCE. It was found in case of dyes **41** (Mi = 1.049; PCE = 2.14), **61** (Mi = 1.041; PCE = 1.3) and **200** (Mi = 1.037; PCE = 2) that with an increase in the value of the mean first ionization potential, there is a significant decrease in the PCE and *vice versa* in case of dyes **44** (Mi = 1.012; PCE = 7.48), **45** (Mi = 1.011; PCE = 6.56) and **100** (Mi = 1.010; PCE = 6.59). The mean first ionization potential is related to polarity of the molecules which plays an essential role to regulate the PCE property of dyes in solar cell. Due to presence of small electronegative atoms, dyes behave like polar molecules. It is known that polarity tends to attain the aggregation of dye molecules on the semiconductor [51]. Thus, for enhancement of the PCE property of dye molecules, the mean first ionization potential of dye molecules should be low.

The positive regression coefficient of the 2D atom pair descriptors like B10[C-S] (presence or absence of carbon and sulfur atoms at the topological distance 10), B07[S-S] (presence or absence of 2 sulfur atoms at the topological distance 7) and F04[O-O] (the frequency of the two oxygen atoms at the topological distance 4) as well as C-022 (which accounts for #CR/R = C = R; where R: any group linked through carbon;

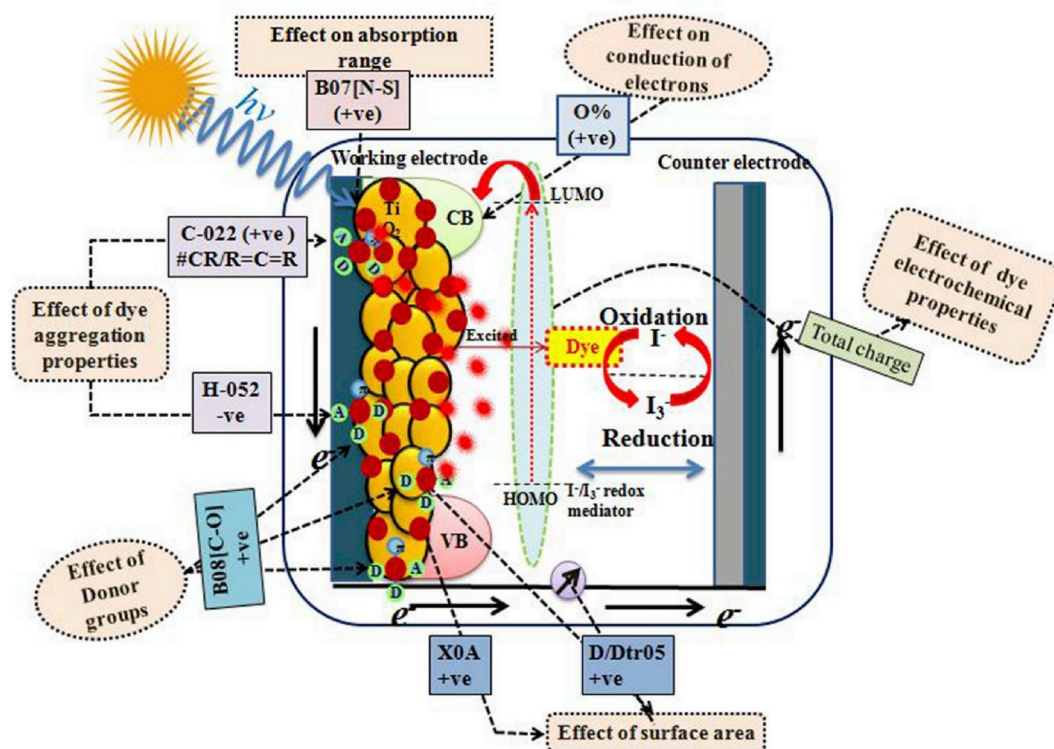


Fig. 6. Contribution of modeled descriptors on controlling PCE values of the Phenothiazine dyes.

X: any electronegative atom O, N, S, P, Se, halogens; = : a double bond; #: a triple bond) indicate that presence of such fragments in the dye molecules may enhance the PCE property in DSSCs as evidenced by the dyes **2** (PCE = 7.5), **3** (PCE = 6.87), **4** (PCE = 7.78) where the numerical value of B10[C-S] descriptor in all three cases is 1; **4** (PCE = 7.78), **78** (PCE = 6.78), **126** (PCE = 7.44), where the numerical value of B07[S-S] descriptor in all these cases is 1; **2** (PCE = 7.5), **51** (PCE = 7.7) and **191** (PCE = 7.1) for which the numerical value of F04[O-O] descriptor in all these cases is 1; and **179** (PCE = 7.44), **180** (PCE = 6.64), **182** (PCE = 6.06) where the numerical value of C-022 in all these examples is 4. On

the other hand, the absence of these fragments may reduce the PCE property as shown in the dyes **1** (PCE = 2.71), **114** (PCE = 1.12) and **189** (PCE = 2.82) for B10[C-S] fragment; **61** (PCE = 1.3), **65** (PCE = 1.3) and **68** (PCE = 1.2) for B07[S-S] descriptor; **27** (PCE = 1.83) and **67** (PCE = 1.9) for F04[O-O] descriptor; and **56** (PCE = 0.73), **57** (PCE = 0.99) and **70** (PCE = 0.6) for C-022 fragment.

On the contrary, 2D atom pair descriptors like B04[O-S] (presence or absence of oxygen and sulfur atoms at topological distance 4), F09[S-S] (frequency of 2 sulfur atoms at topological distance 9) and F05[C-O] (frequency of carbon and oxygen atoms at topological distance 5)

### Box 3

#### IM1

$$\text{PCE} = 2.982 + 0.907 \times \text{nCrq} + 0.992 \times \text{B07[N-N]} - 1.520 \times \text{B04[S-S]} - 0.733 \times \text{NsssN} + 0.272 \times \text{F04[C-N]} - 2.060 \times \text{B06[N-N]} + 1.382 \times \text{B05[O-S]} - 6.394 \times \text{SaaaC} + 0.089 \times \text{F10[C-N]} - 0.825 \times \text{B05[S-S]} + 0.930 \times \text{F07[N-S]} + 0.649 \times \text{B09[O-S]} + 0.642 \times \text{F07[N-O]}$$

#### IM2

$$\text{PCE} = 3.004 + 1.042 \times \text{nCrq} + 1.224 \times \text{B07[N-N]} + 0.268 \times \text{F04[C-N]} - 1.206 \times \text{F04[S-S]} - 2.200 \times \text{B06[N-N]} + 1.327 \times \text{B05[O-S]} - 6.174 \times \text{SaaaC} + -0.097 \times \text{F10[C-N]} - 0.949 \times \text{B05[S-S]} + 0.639 \times \text{F07[N-O]} + 0.997 \times \text{F07[N-S]} + 0.482 \times \text{B09[O-S]}$$

#### IM3

$$\text{PCE} = 3.119 + 1.045 \times \text{nCrq} + 0.866 \times \text{F07[N-N]} - 0.291 \times \text{B04[S-S]} + 0.278 \times \text{F04[C-N]} - 6.600 \times \text{SaaaC} - 0.758 \times \text{NsssN} - 1.162 \times \text{F04[S-S]} + 1.404 \times \text{B05[O-S]} - 1.968 \times \text{B06[N-N]} - 0.111 \times \text{F10[C-N]} + 1.006 \times \text{F07[N-S]} - 1.014 \times \text{B05[S-S]} + 0.734 \times \text{F07[N-O]}$$

#### IM4

$$\text{PCE} = 3.161 + 1.123 \times \text{B07[N-N]} - 0.950 \times \text{nCrq} - 1.549 \times \text{B04[S-S]} - 0.039 \times \text{nConj} + 0.276 \times \text{F04[C-N]} - 0.697 \times \text{NsssN} - 2.059 \times \text{B06[N-N]} + 1.288 \times \text{B05[O-S]} - 6.516 \times \text{SaaaC} - 0.095 \times \text{F10[C-N]} + 1.018 \times \text{F07[N-S]} + 0.644 \times \text{F07[N-O]} - 0.720 \times \text{B05[S-S]}$$

#### IM5

$$\text{PCE} = 3.119 + 0.922 \times \text{nCrq} + 1.123 \times \text{B07[N-N]} - 1.101 \times \text{B02[N-O]} - 0.849 \times \text{B04[S-S]} + 0.287 \times \text{F04[C-N]} - 0.576 \times \text{NsssN} + 1.119 \times \text{B05[O-S]} - 6.387 \times \text{SaaaC} - 2.143 \times \text{B06[N-N]} - 0.707 \times \text{F10[O-S]} - 0.124 \times \text{F10[C-N]} + 0.665 \times \text{F07[N-O]} + 1.056 \times \text{F07[N-S]}$$

contribute negatively towards the PCE property of dyes in DSSC. The negative regression coefficients of these descriptors indicate that presence of such features in the structures of dye molecules may reduce the PCE property as observed in the dyes **65** (PCE = 1.3), **68** (PCE = 1.2), **71** (PCE = 1.3) for which the numerical value of B04[O-S] is 1 in all cases; dye **206** (PCE = 1.5) for which the F09[S-S] value is 2; **155** (PCE = 0.12) and **157** (PCE = 0.93) with F05[C-O] descriptor values of 20 and 19, respectively. On the other hand, absence of such features in the dyes may enhance the PCE property in DSSCs as shown in the dyes like **51** (PCE = 7.7), **132** (PCE = 6.13), **180** (PCE = 6.64) for B04[O-S] and **141** (PCE = 7.87) for F09[S-S]. Dyes **126** (PCE = 6.9) and **192** (PCE = 7.44) showed quite higher PCE due to much lower descriptor values for F05 [C-O] than other dyes *i.e.* 2.

The mechanistic interpretation of the phenothiazine dyes from all the models is schematically portrayed in Fig. 6. The scatter plots of observed vs. predicted PCE property related to the Phenothiazine dyes for all the PLS models are depicted in Fig. S8 in Supplementary material.

### 3.3. Dataset 3: modeling of PCE property of indoline dyes

Significant five PLS models are reported in Box 3. The descriptors appearing in the models are explained below with the most feasible mechanistic interpretation towards the PCE property of dyes in DSSCs.

The positive regression coefficients of the 2D atom pair descriptors like F04[C-N] (frequency of carbon and nitrogen atoms at topological distance 4), F07[N-S] (frequency of nitrogen and sulfur atoms at topological distance 7) and F07[N-O] (frequency of nitrogen and oxygen atoms at topological distance 7) indicate that presence of these fragments in the indoline dyes may enhance the PCE as shown in dyes **13**

(F04[C-N] = 21; PCE = 8.43), **18** (F04[C-N] = 18; PCE = 7.28) and **24** (F04[C-N] = 18; PCE = 9.2) for F04[C-N]; **8** (F07[N-S] = 1; PCE = 7.12), **107** (F07[N-S] = 2; PCE = 7.63) and **146** (F07[N-S] = 4; PCE = 6.71) for F07[N-S]; and **43** (F07[N-O] = 4; PCE = 5.4), **112** (F07[N-O] = 3; PCE = 6.51) and **115** (F07[N-O] = 4; PCE = 5.93) for F07[N-O] and *vice versa* in case of dyes **147** (F04[C-N] = 5, F07[N-S] = 0, F04[N-O] = 0; PCE = 1.8), **149** (F04[C-N] = 4, F07[N-S] = 0, F04[N-O] = 0; PCE = 1.92) and **164** (F04[C-N] = 3, F07[N-S] = 0, F04[N-O] = 0; PCE = 2.08). The dyes containing donor group and groups with non-planar structure are very important for the PCE property. It was already reported that the above mentioned descriptors are present as a part of dye donors with non-planar structures of the Indoline dyes [54, 57].

The other 2D atom pair descriptors F10[O-S] (frequency of oxygen and sulfur atoms at the topological distance 10) and F10[C-N] (frequency of carbon and nitrogen atoms at the topological distance 10) contributed negatively towards the PCE of indoline dyes. These descriptors actually represent bulk of dye molecules which may weaken the interactions between the semiconductor and the dye molecules due to steric hindrance followed by restriction of the transfer of electrons from the dye molecules to the semiconductor [58]. Therefore, the dyes bearing such fragments may reduce the PCE property of dyes in DSSC as shown in the dyes **14** (F10[O-S] = 2; PCE = 3.85), **53** (F10[O-S] = 2; PCE = 2.96), and **150** (F10[O-S] = 2; PCE = 2.1), for F10[O-S]; and **34** (F10[C-N] = 6; PCE = 2.7), **35** (F10[C-N] = 5; PCE = 1.2) and **55** (F10[C-N] = 7; PCE = 3.9) for F10[C-N] and *vice versa* in case of dyes **44** (F10[O-S] = 0, F10[C-N] = 3; PCE = 6.96), **59** (F10[O-S] = 0, F10[C-N] = 2; PCE = 8.34) and **94** (F10[O-S] = 0, F10[C-N] = 2; PCE = 8.42).

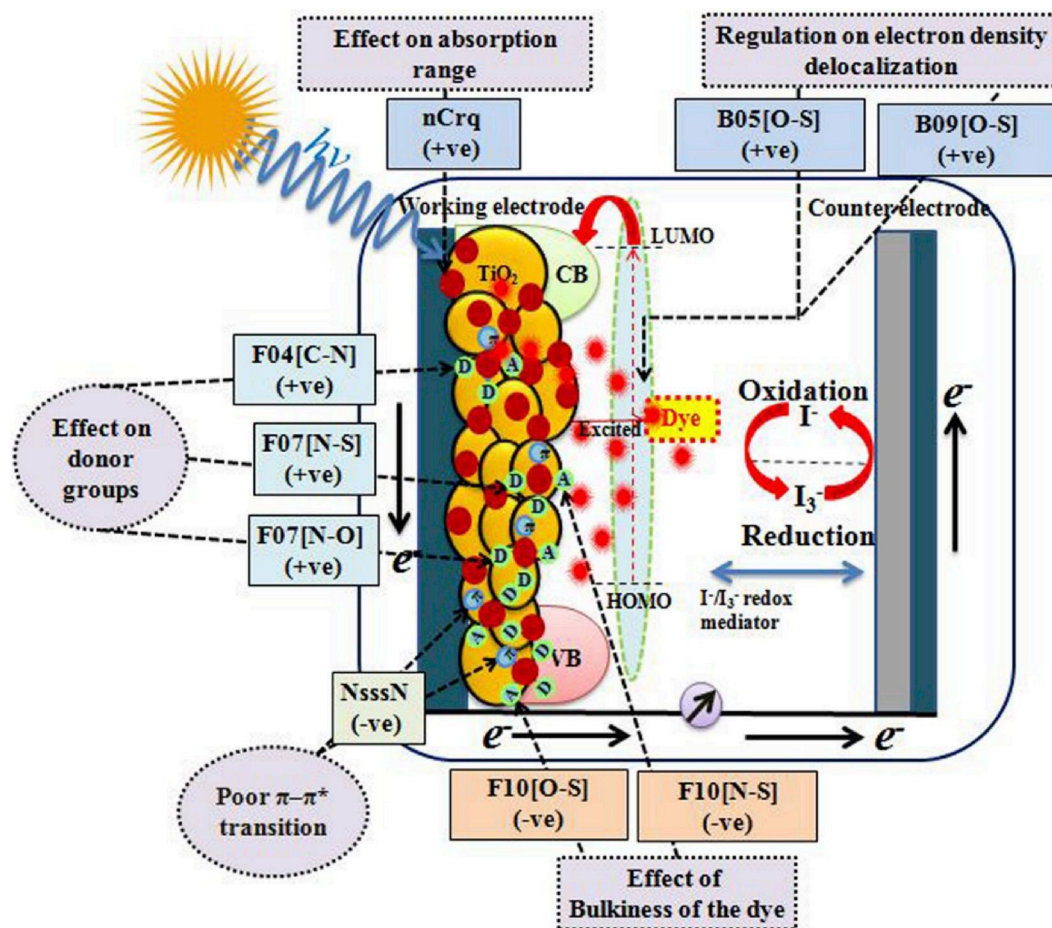


Fig. 7. Contribution of features on controlling PCE values of the Indoline dyes.



**Box 4**

IM1

$$\text{PCE} = -16.058 + 0.560 \times \text{F06}[\text{O}-\text{X}] + 0.296 \times \text{F07}[\text{C}-\text{X}] - 0.421 \times \text{O}\% + 224.846 \times \text{X4A} - 0.067 \times \text{F08}[\text{C}-\text{O}] - 2.395 \times \text{B04}[\text{N}-\text{N}] + 0.359 \times \text{F10}[\text{N}-\text{O}] - 2.179 \times (\text{N}-072) - 1.155 \times \text{F03}[\text{N}-\text{X}] + 2.267 \times \text{B05}[\text{N}-\text{S}] - 0.191 \times \text{SdsCH} + 0.887 \times \text{F03}[\text{N}-\text{O}]$$

IM2

$$\text{PCE} = -16.139 + 0.294 \times \text{F07}[\text{C}-\text{X}] + 0.556 \times \text{F06}[\text{O}-\text{X}] - 0.407 \times \text{O}\% + 224.846 \times \text{X4A} - 0.071 \times \text{F08}[\text{C}-\text{O}] - 2.434 \times \text{B04}[\text{N}-\text{N}] + 0.347 \times \text{F10}[\text{N}-\text{O}] - 2.130 \times (\text{N}-072) - 1.126 \times \text{F03}[\text{N}-\text{X}] + 2.228 \times \text{B05}[\text{N}-\text{S}] - 0.189 \times \text{SdsCH} + 1.810 \times \text{B03}[\text{N}-\text{O}]$$

IM3

$$\text{PCE} = -14.909 - 0.429 \times \text{O}\% + 0.291 \times \text{F07}[\text{C}-\text{X}] + 0.574 \times \text{F06}[\text{O}-\text{X}] + 191.057 \times \text{X4A} - 0.066 \times \text{F08}[\text{C}-\text{O}] - 2.434 \times \text{B04}[\text{N}-\text{N}] + 0.403 \times \text{F10}[\text{N}-\text{O}] - 2.079 \times (\text{N}-071) + 2.651 \times \text{B05}[\text{N}-\text{S}] - 2.079 \times (\text{N}-072) - 0.216 \times \text{SdsCH} - 1.196 \times \text{F03}[\text{N}-\text{X}] + 0.988 \times \text{F03}[\text{N}-\text{O}]$$


IM4

$$\text{PCE} = -15.101 - 0.413 \times \text{O}\% + 0.288 \times \text{F07}[\text{C}-\text{X}] + 0.571 \times \text{F06}[\text{O}-\text{X}] + 193.418 \times \text{X4A} - 0.070 \times \text{F08}[\text{C}-\text{O}] + 0.390 \times \text{F10}[\text{N}-\text{O}] - 1.157 \times (\text{N}-071) + 2.619 \times \text{B05}[\text{N}-\text{S}] - 0.215 \times \text{SdsCH} - 2.045 \times (\text{N}-072) - 1.173 \times \text{F03}[\text{N}-\text{X}] + 1.946 \times \text{B03}[\text{N}-\text{O}]$$

IM5

$$\text{PCE} = -14.102 + 0.362 \times \text{F07}[\text{C}-\text{X}] + 202.714 \times \text{X4A} - 0.483 \times \text{O}\% + 1.237 \times (\text{N}-071) - 2.755 \times \text{B04}[\text{N}-\text{N}] + 2.322 \times \text{B05}[\text{N}-\text{S}] - 1.109 \times \text{F03}[\text{N}-\text{X}] + 0.287 \times \text{F10}[\text{N}-\text{O}] - 0.177 \times \text{SdsCH} - 1.987 \times (\text{N}-072) + 0.802 \times \text{F03}[\text{N}-\text{O}] - 1.101 \times \text{nR08}$$

The functional group count descriptor nCrq indicates the number of ring quaternary carbons with  $\text{sp}^3$  hybridization, and it has a positive contribution towards the PCE. The presence of ring quaternary carbon with  $\text{sp}^3$  hybridization is essential for tunable absorption properties and it produces high molar extinction coefficients leads to improve the energy level reactions in the solar cell [59]. Thus, presence of higher number of  $\text{sp}^3$  hybridized quaternary carbon atom in dyes may enhance the PCE as observed in dyes like **126** (nCrq = 1; PCE = 6.9), **144** (nCrq = 3; PCE = 8.78) and **157** (nCrq = 3; PCE = 7.08). In contrary, the absence of such type of fragment in the dyes may reduce the PCE as evidenced by the dyes **32** (nCrq = 0; PCE = 1.48), **93** (nCrq = 0; PCE = 0.35), and **129** (nCrq = 0; PCE = 1.48).

The count of atom-type E-State descriptor NsssN states the number of atoms of type sssN () which has a negative contribution towards the PCE. Due to the presence of such fragments, the dyes experience poor  $\pi-\pi^*$  transition (the fragment is less reactive than imines) which results in slow energy cascade mechanism [22]. Thus, the dyes containing such fragments may have lower PCE property in DSSCs as evidenced by dyes **27** (NsssN = 4; PCE = 2.12), **99** (NsssN = 3; PCE = 1.42) and **169** (NsssN = 3; PCE = 1.71) and *vice versa* in case of dyes **77** (NsssN = 0; PCE = 6.86), **78** (NsssN = 0; PCE = 7.99) and **146** (NsssN = 0; PCE = 6.71).

The positive coefficient of other 2D atom pair descriptors B05[O-S] (presence or absence of oxygen and sulfur atoms at topological distance 5) and B09[O-S] (presence or absence of oxygen and sulfur atoms at topological distance 9) indicated that presence of these fragments in dyes may increase the PCE property in DSSCs. In the dye system, oxygen and sulfur atoms regulate the electron density delocalization which is favorable for the  $\pi$ -bond conjugation. As a result, the molar extinction coefficient of the dye is enhanced which leads to the bathochromic shift of the absorption spectrum [45]. Therefore, the dyes containing these fragments may show good PCE property in DSSCs as evident by the dyes **143** (B05[O-S] = 1; PCE = 7.25), **144** (B05[O-S] = 1; PCE = 8.78) and **145** (B05[O-S] = 1; PCE = 7.4) for B05[O-S]; and **21** (B09[O-S] = 1; PCE = 6.12), **78** (B09[O-S] = 1; PCE = 7.99) and **131** (B09[O-S] = 1; PCE = 6.11) for B09[O-S] and *vice versa* in case of dyes **30** (B05[O-S], B09[O-S] = 0; PCE = 0.77), **32** (B05[O-S], B09[O-S] = 0; PCE = 0.63) and **35** (B05[O-S], B09[O-S] = 0; PCE = 1.2).

The 2D atom pair descriptors F07[N-N] (frequency of 2 nitrogen atoms at the topological distance 7), B07[N-N] (presence or absence of 2

nitrogen atoms at topological distance 7) and the functional group count descriptor nCconj (the number of non-aromatic conjugated carbon with  $\text{sp}^2$  hybridization) contributed positively towards the PCE which indicates that the PCE values increase with an increase in the numerical value of these descriptors as shown in dyes **8** (B07[N-N] = 2; PCE = 7.12), **115** (B07[N-N] = 2; PCE = 5.93) for F07[N-N]; **135** (B07[N-N] = 1; PCE = 8.38), **141** (B07[N-N] = 1; PCE = 8.61) for B07[N-N]; and **13** (nCconj = 13; PCE = 8.43), **18** (nCconj = 13; PCE = 7.28) and *vice versa* in case of dyes **32** (F07[N-N] = 0, B07[N-N] = 0, nCconj = 1; PCE = 0.63), **93** (F07[N-N] = 0, B07[N-N] = 0, nCconj = 3; PCE = 0.35) and **108** (F07[N-N] = 0, B07[N-N] = 0, nCconj = 3; PCE = 0.046).

The 2D atom pair descriptor B06[N-N] represents the presence or absence of 2 nitrogen atoms at the topological distance 6. B04[S-S] means presence or absence of 2 sulfur atoms at the topological distance 4 and F04[S-S] stands for frequency of 2 sulfur atoms at the topological distance 4, B02[N-O] states presence or absence of nitrogen and oxygen atoms at the topological distance 2 and the atom centered fragment SaaaC represents sum of aromatic carbons ( $-\text{C}(-)-$ ) ( $-$ represents an aromatic bond). The negative regression coefficients of these descriptors indicate that an increase in the numerical values of the descriptors may reduce the PCE values as observed in the dyes **93** (B06[N-N] = 1, B04[S-S] = 1 & F04[S-S] = 1, B02[N-O] = 1; PCE = 0.35), **108** (B06[N-N] = 1, B04[S-S] = 1 & F04[S-S] = 1, B02[N-O] = 1; PCE = 0.046), **30** (SaaaC = 0.697; PCE = 0.77) and **32** (SaaaC = 0.857; PCE = 0.63). In contrary, compounds having no such fragments may show higher PCE values as observed in case of dyes **18** (B06[N-N] = 0, B04[S-S] = 0, F04[S-S] = 0, B02[N-O] = 0; PCE = 7.08), **157** (B06[N-N] = 0, B04[S-S] = 0, F04[S-S] = 0, B02[N-O] = 0; PCE = 7.79), **13** (PCE = 8.43) and **144** (PCE = 8.78).

The mechanistic interpretation from all models is schematically portrayed in Fig. 7 for Indoline dyes. The scatter plots of observed vs. predicted PCE property related to the indoline dyes for all the PLS models are depicted in Fig. S9 in Supplementary material.

### 3.4. Dataset 4: modeling of PCE property of porphyrin dyes

The modeled descriptors for porphyrin dyes obtained from the five PLS models are illustrated in Box 4. The best possible mechanistic interpretation of the descriptors is discussed below with the examples of studied dyes.

The 2D atom pair descriptors F10[N-O] (frequency of N-O at

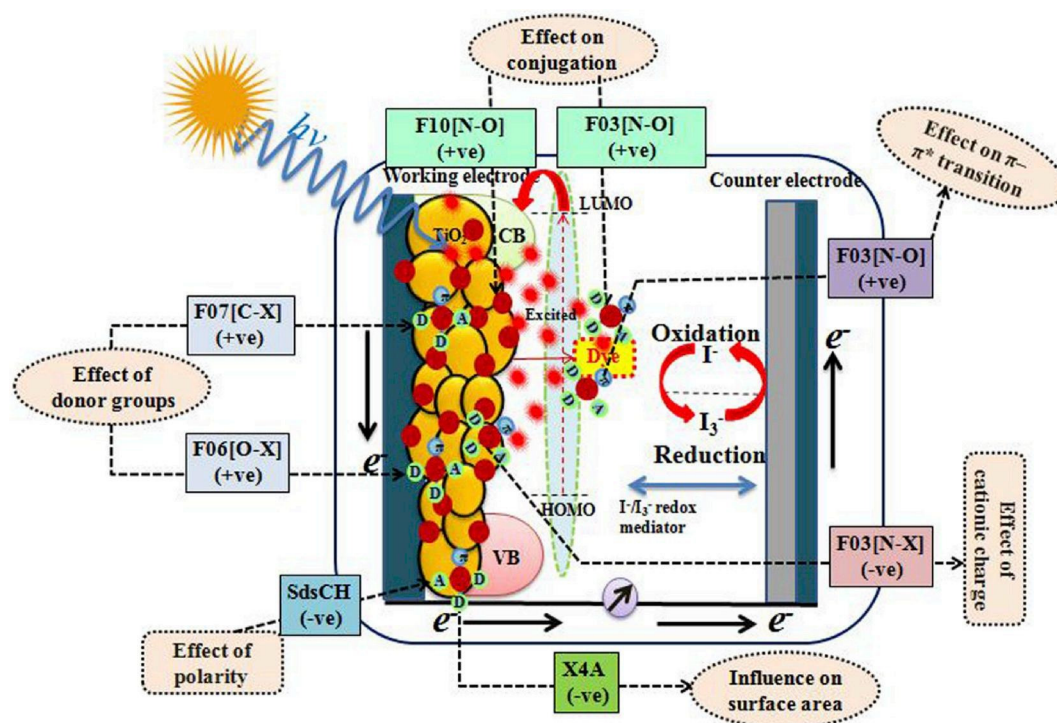


Fig. 8. Contribution of significant features on controlling PCE values of the porphyrin dyes.

topological distance 10) and F03[N-O] (frequency of N-O at topological distance 3) contributed positively towards the PCE. Thus, presence of such fragments in the dyes may enhance the PCE as showed in dyes **37** (F10[N-O] = 6; PCE = 9.1), **38** (F10[N-O] = 8; PCE = 7.2), **39** (F10[N-O] = 8; PCE = 6.9) for F10[N-O]; **34** (F03[N-O] = 2; PCE = 8.3), **208** (F03[N-O] = 1; PCE = 7.94), **213** (F03[N-O] = 1; PCE = 8.6) for F03[N-O], while absence of these fragments may reduce the PCE property as verified by the dyes **113** (PCE = 2.46), **265** (PCE = 0.63) and **271** (PCE = 0.38). Presence of nitrogen and oxygen atoms has a contribution to the special stability (conjugation) and allows the smaller HOMO-LUMO gap, followed by the red shift of the absorption spectrum [60]. Thus, the conduction valence edge level increases, and in result the PCE values of DSSCs increases.

The 2D atom pair descriptors F06[O-X] and F07[C-X] denote the frequency of carbon and heavy metal (X) at topological distance 6 and 7, respectively. They contribute positively towards the PCE. The descriptor F07[C-X] signifies the carbon is a part of meso aryl substituted portion of Zinc porphyrin which will act as a donor in the dye [61]. On the other hand, in case of F06[O-X], the oxygen atom is a part of the long-chain alkoxy group which impairs interfacial back electron transfer reaction [62]. Thus, electron donating ability and presence of electron back transfer groups may increase the PCE property as shown in the dye molecules **40** (F07[C-X] = 12, F06[O-X] = 4; PCE = 8.6), **48** (F07[C-X] = 12, F06[O-X] = 4; PCE = 8.42) and **99** (F07[C-X] = 10, F06[O-X] = 4; PCE = 9.19) and *vice versa* in case of dyes **63** (F07[C-X] = 0, F06[O-X] = 0; PCE = 0.42), **102** (F07[C-X] = 0, F06[O-X] = 0; PCE = 0.3) and **123** (F07[C-X] = 0, F06[O-X] = 0; PCE = 0.11).

The atom type E-state index SdsCH defines the E-state atom index of C atom of the fragment dsCH (=CH-) which contributes negatively to the PCE property. Therefore, presence of this fragment in the dyes decreases the PCE property as observed in dyes **96** (SdsCH = 25.43; PCE = 2.3), **263** (SdsCH = 18.19; PCE = 1.64) and **268** (SdsCH = 19.21; PCE = 2.37) and *vice versa* in case of dyes **141** (SdsCH = 1.41; PCE = 8), **217** (SdsCH = 1.39; PCE = 7.88) and **218** (SdsCH = 1.39; PCE = 8.14). Presence of this non-polar group makes the negative shift in the solvatochromic properties (ability of chemical substance to change color due to a change in polarity) of the dye. Thus, the dyes cannot adhere properly to the

semiconductor which results in a negative effect on absorption and stability of the dye [63].

Another significant descriptor X4A indicates average connectivity index of order 4; it encodes the 'χ' value across four bonds which can be calculated on the basis of Kier and Hall's connectivity index [64]. It contributes positively towards the PCE. This indicates that the PCE property of dyes increases with an increase in the numerical value of this descriptor as shown in dyes **81** (X4A = 0.1123; PCE = 10.17), **196** (X4A = 0.1123; PCE = 9.25) and **201** (X4A = 0.1145; PCE = 10.24) and *vice versa* in case of dyes **42** (X4A = 0.097; PCE = 0.6), **45** (X4A = 0.096; PCE = 0.02) and **105** (X4A = 0.096; PCE = 1.1). This descriptor is related to surface area of dyes which is directly related to light-harvesting capability which could be achieved maximum when the surface area of the dyes is large [53].

The 2D atom pair descriptor B03[N-O] indicates the presence or absence of nitrogen and oxygen atoms at the topological distance 3 which offers a positive effect to the PCE property as evidenced by dyes **34** (PCE = 8.3), **213** (PCE = 8.6) and **214** (PCE = 8.7) due to their descriptor values being equal to 1 in all cases and *vice versa* (descriptor value zero) for dyes **44** (PCE = 0.0013), **45** (PCE = 0.02) and **286** (PCE = 0.03) in absence of this fragment. This fragment may represent the conjugation units to 'π' system of the dye, which engenders a lower internal resistance to the transport of positive charges. Thus, the conjugated π-system may enhance the PCE property of dye molecules in DSSC [65].

The 2D atom pair descriptor F03[N-X] denotes frequency of the nitrogens and heavy metal atoms (Zn) at the topological distance 3 which contributes negatively towards the PCE property as indicated by bearing dyes like **252** (F03[N-X] = 4; PCE = 0.73), **253** (F03[N-X] = 4; PCE = 1.54) and **292** (F03[N-X] = 4; PCE = 0.92) which have lower PCE values. Again, dyes not having this feature showed higher range of PCE as evidenced by the dyes **213** (F03[N-X] = 0; PCE = 8.6), **214** (F03[N-X] = 0; PCE = 8.7) and **216** (F03[N-X] = 0; PCE = 9.5). The presence of the nitrogen atom at topological distance 3 from Zinc in the porphyrin moiety has a positive influence on the cationic charge of the dye endowing the aggregation resulting in low PCE values [66].

The positive regression coefficients of 2D atom pair descriptor B05

## Box 5

IM1

$$\text{PCE} = -1.672 + 1.383 \times \text{nRCN} + 0.979 \times \text{F08}[\text{N} - \text{S}] + 2.615 \times \text{nArNR2} - 3.358 \times \text{B09}[\text{S} - \text{S}] + 0.261 \times \text{nCconj}$$

IM2

$$\text{PCE} = -1.628 + 1.455 \times \text{nRCN} + 2.647 \times \text{nArNR2} - 3.508 \times \text{B09}[\text{S} - \text{S}] + 1.115 \times \text{B08}[\text{N} - \text{S}] + 0.249 \times \text{nCconj}$$

IM3

$$\text{PCE} = -1.953 + 1.265 \times \text{nRCN} + 0.323 \times (\text{C} - 034) + 2.658 \times \text{nArNR2} - 3.354 \times \text{B09}[\text{S} - \text{S}] + 0.300 \times \text{nCconj}$$

IM4

$$\text{PCE} = -1.574 + 1.398 \times \text{nRCN} + 0.627 \times \text{nThiophenes} + 2.649 \times \text{nArNR2} - 3.229 \times \text{B09}[\text{S} - \text{S}] + 0.230 \times \text{nCconj}$$

IM5

$$\text{PCE} = 0.082 + 1.151 \times \text{nR} = \text{Ct} + 0.539 \times (\text{C} - 034) - 0.098 \times \text{T}(\text{S..S}) - 0.946 \times \text{nR}\#\text{C} - + 0.539 \times (\text{C} - 040)$$

[N-S] which signifies the presence or absence of the nitrogen and sulfur atoms at topological distance 5 and the atom centered fragment descriptor N-071 (Ar-NAI2; where, Al and Ar: aliphatic and aromatic groups, respectively) indicate that the presence of these fragments is influential for the PCE property of dyes in DSSC as observed for the dyes **217** (PCE = 7.88) and **218** (PCE = 8.14) for which B05[N-S] value is 1 for both cases; and **178** (PCE = 9.73) and **180** (PCE = 9.51) for which N-071 descriptor value is 1 for both cases. In contrast, the dyes having no such fragments showed poor PCE property in DSSC as observed in case of dyes like **92** (PCE = 0.65), **252** (PCE = 0.73) and **292** (PCE = 0.92).

The 2D atom pair descriptors B04[N-N] (presence or absence of 2 nitrogens at the topological distance 4), F08[C-O] (frequency of carbon and oxygen atoms at the topological distance 8) and the atom centered fragment N-072(RCO-N</>N-X = X), the constitutional descriptor O% (percentage of oxygen atoms) as well as the ring descriptor nR08 (number of 8 membered rings) contribute negatively towards the PCE property as suggested by their negative regression coefficients. As the

numerical values of these descriptors increase, the PCE property of the dyes will decrease. For example, the dyes **292** (PCE = 0.92) and **294** (PCE = 0.133) for B04[N-N] descriptor where the numerical value is 1 for both cases; **252** (PCE = 0.73) and **120** (PCE = 2.48) for which the numerical values of F08[C-O] are 60 and 46, respectively; **120** (N-072 = 2, nR08 = 1; PCE = 2.48) and **121** (N-072 = 2, nR08 = 1; PCE = 2.58) for N-072 and nR08 descriptors; **44** (PCE = 0.6) and **45** (PCE = 0.02) for O% descriptor with the numerical value of 9.41 in both cases have lower PCE values and *vice versa* in case of dyes **217** (PCE = 7.88) and **218** (PCE = 8.14) where B04[N-N] is absent; **188** (PCE = 8.1) and **198** (PCE = 8.77) with descriptor value being 2 for F08[C-O] in both cases; **214** (PCE = 8.7) and **216** (PCE = 9.5) in absence of both features N-072 and nR08; **198** (PCE = 8.77) and **205** (PCE = 8.26) for O% where the descriptor values are 1.09 and 1.38, respectively.

The mechanistic interpretation of the models for porphyrin dyes is schematically portrayed in Fig. 8. The scatter plots of observed vs. predicted PCE property related to the Porphyrin dyes for all the PLS

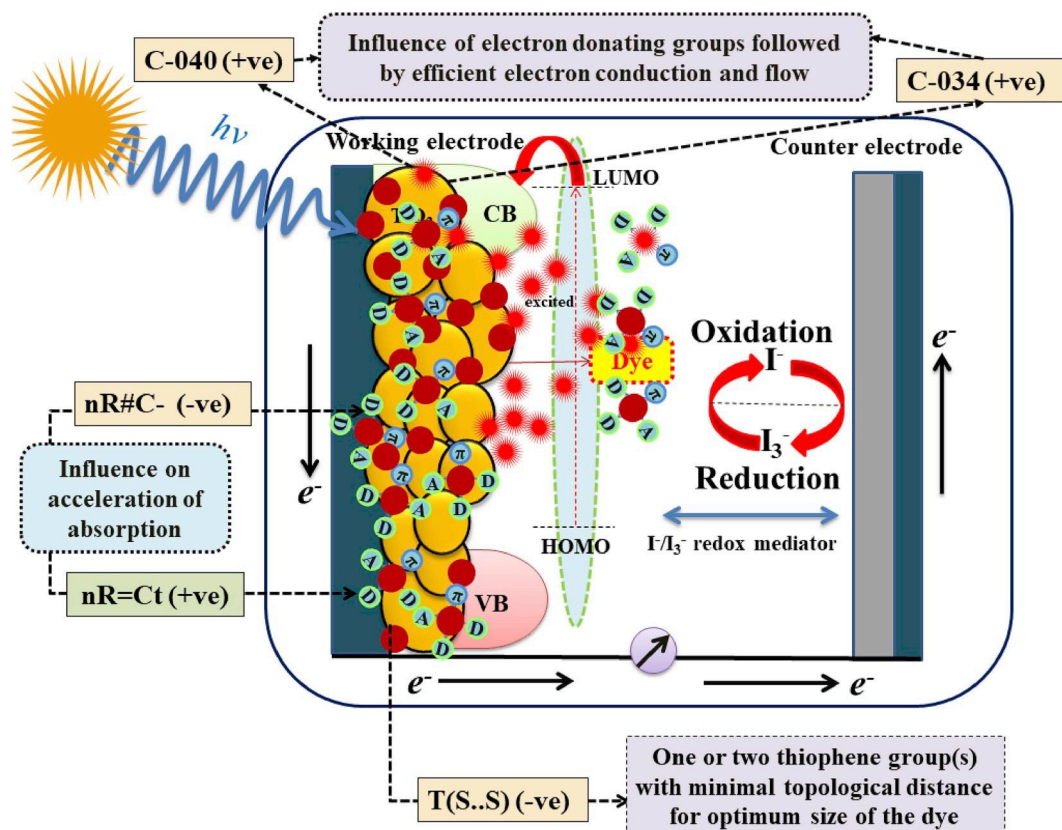


Fig. 9. Contribution of modeled descriptors on controlling PCE values of the Coumarin dyes.



**Box 6****IM 1**

$$\text{PCE} = 0.541 + 0.019 \times \text{F06}[\text{C} - \text{C}] - 0.154 \times \text{NaasC} + 1.808 \times \text{B04}[\text{N} - \text{O}] + 1.409 \times \text{F08}[\text{O} - \text{O}] + 1.339 \times \text{B02}[\text{C} - \text{S}] + 1.177 \times \text{F06}[\text{N} - \text{O}] + 0.918 \times \text{B08}[\text{O} - \text{S}] - 0.266 \times \text{N\%} + 0.143 \times \text{F04}[\text{C} - \text{N}] + 0.601 \times \text{nR10} - 1.445 \times \text{F06}[\text{N} - \text{N}]$$

**IM2**

$$\text{PCE} = 0.177 + 0.019 \times \text{F06}[\text{C} - \text{C}] - 0.125 \times \text{NaasC} + 1.933 \times \text{B04}[\text{N} - \text{O}] + 1.263 \times \text{F08}[\text{O} - \text{O}] + 1.261 \times \text{B10}[\text{C} - \text{S}] + 1.072 \times \text{F06}[\text{N} - \text{O}] + 0.949 \times \text{B08}[\text{O} - \text{S}] - 0.202 \times \text{N\%} + 0.129 \times \text{F04}[\text{C} - \text{N}] + 0.714 \times \text{nR10} - 1.444 \times \text{F06}[\text{N} - \text{N}]$$

**IM3**

$$\text{PCE} = -0.415 + 0.021 \times \text{F06}[\text{C} - \text{C}] - 0.141 \times \text{NaasC} + 1.734 \times \text{B04}[\text{N} - \text{O}] + 1.377 \times \text{F08}[\text{O} - \text{O}] + 1.009 \times \text{B02}[\text{C} - \text{S}] + 1.108 \times \text{F06}[\text{N} - \text{O}] + 0.638 \times \text{B06}[\text{N} - \text{S}] + 1.007 \times \text{B08}[\text{O} - \text{S}] + 0.732 \times \text{nR10} + 0.114 \times \text{F04}[\text{C} - \text{N}] - 1.588 \times \text{F06}[\text{N} - \text{N}]$$

**IM4**

$$\text{PCE} = 0.576 + 0.016 \times \text{F06}[\text{C} - \text{C}] - 0.098 \times \text{NaasC} + 2.044 \times \text{B04}[\text{N} - \text{O}] + 1.652 \times \text{B02}[\text{C} - \text{S}] + 1.614 \times \text{F08}[\text{O} - \text{O}] + 0.933 \times \text{F06}[\text{N} - \text{O}] - 0.295 \times \text{N\%} + 0.108 \times \text{F04}[\text{C} - \text{N}] + 0.694 \times \text{nR10} - 0.029 \times \text{F06}[\text{O} - \text{S}] - 1.346 \times \text{F06}[\text{N} - \text{N}]$$

**IM5**

$$\text{PCE} = 0.307 + 0.020 \times \text{F06}[\text{C} - \text{C}] - 0.130 \times \text{NaasC} + 1.804 \times \text{B04}[\text{N} - \text{O}] + 2.442 \times \text{F08}[\text{O} - \text{O}] + 1.206 \times \text{B10}[\text{C} - \text{S}] + 0.853 \times \text{B08}[\text{O} - \text{S}] - 0.189 \times \text{N\%} + 0.133 \times \text{F04}[\text{C} - \text{N}] + 0.670 \times \text{nR10} - 1.547 \times \text{F06}[\text{N} - \text{N}] - 0.546 \times \text{B04}[\text{O} - \text{S}]$$

models are depicted in Fig. S10 in Supplementary Materials.

### 3.5. Data set 5: modeling of PCE property of coumarin dyes

The identified and extracted vital features obtained from the five PLS models for coumarin dyes are reported in Box 5. The mechanistic interpretation of all the descriptors with suitable examples of the studied dyes is discussed below to demonstrate how the features are affecting the PCE property.

Unlike other chemical classes, here we are not discussing descriptor interpretations for consensus models. IM5 is the best model for coumarin dyes, and the modeled descriptors only are discussed below. The functional group count descriptors nR#C- (number of non-terminal carbon with “sp” hybridization) and nR = Ct (number of aliphatic tertiary carbon with “sp<sup>2</sup>” hybridization) had negative and positive contributions respectively towards PCE property of coumarin dyes. Absence of non-terminal “sp” hybridized C atom and presence of tertiary aliphatic “sp<sup>2</sup>” hybridized C atoms in the dyes enhance the frequency of the S-character (S means strong absorption) which is important to the enhancement of PCE values [67].

The atom centered fragments C-034 (R–CR..X) and C-040 (R–C(=X)–X/R–C#X/X = C = X) (where, R: any group linked through carbon; X: any electronegative atom O, N, S, P, Se, halogens; #: a triple bond; –: an aromatic bond as in benzene or delocalized bonds such as the N–O bond in a nitro group; ..: aromatic single bonds as the C–N bond in pyrrole) contribute positively to the PCE. Thus, the presence of such fragments in the dyes may enhance the PCE property in DSSC as observed in case of dyes **19** (F08[N–S] = 2, C-034 = 3, C-040 = 4; PCE = 7.4) and **32** (F08[N–S] = 1, C-034 = 4, C-040 = 4; PCE = 6.5) and *vice versa* in case of dyes **7** (F08[N–S] = 0, C-034 = 2, C-040 = 1; PCE = 1.1) and **24** (F08[N–S] = 0, C-034 = 2, C-040 = 0; PCE = 1.04).

The 2D atom pair descriptor T(S..S) means the sum of topological distance between two sulfur atoms (where, .. signifies aromatic single bonds), and it shows a negative contribution to the PCE of Coumarin dyes. It has been found that with the an increase in the numerical value of this descriptor, the PCE property of dyes decreases, as clearly observed in case of dyes **1** (T(S..S) = 52; PCE = 1.39) and **3** (T(S..S) = 28; PCE = 1.77), while in case of dyes **29** (T(S..S) = 3; PCE = 6.07) and **35** (T(S..S) = 0; PCE = 6.20), the PCE values are increased due to absence of such fragment.

The mechanistic interpretation of the descriptors for coumarin dyes from all the models is schematically portrayed in Fig. 9. The scatter plots of observed vs. predicted PCE property related to the coumarin dyes for all the PLS models are depicted in Fig. S11 in Supplementary Material.

### 3.6. Data set 6: modeling of PCE property of carbazole dyes

The modeled descriptors obtained from the five PLS models for carbazole dyes are reported in Box 6. The mechanistic interpretation of all the descriptors is discussed below with suitable examples.

The 2D atom pair descriptor F06[C–C] denotes the frequency of two carbon atoms at the topological distance 6 with a positive effect towards the PCE. Thus, dyes bearing this fragment like **99** (F06[C–C] = 149; PCE = 7.58), **101** (F06[C–C] = 139; PCE = 8.09) and **130** (F06[C–C] = 207; PCE = 9.8) showed higher range of PCE property in DSSC. Again, dyes having lower frequency of this fragment showed lower range of PCE value as evidenced by the dyes **47** (F06[C–C] = 14; PCE = 2.43), **97** (F06[C–C] = 0; PCE = 0.0538) and **98** (F06[C–C] = 0; PCE = 0.0387). Presence of this descriptor signifies the importance of alkyl linear chains in the dyes which may enhance the electron transfer (these chains improve the surface protection which will facilitate electron injection [68] from donor to the acceptor resulting in generation of the charge separated species followed by an increase in the PCE value [69]).

Another 2D atom pair descriptor F08[O–O] represents the frequency of two oxygen atoms at topological distance 8, which contributes positively towards the PCE. We can see in the dyes **132** (F08[O–O] = 3; PCE = 12.5) and **133** (F08[O–O] = 2; PCE = 9.32), the PCE property is high due to the higher numerical value of this descriptor, and the opposite is observed in case of dyes **62** (PCE = 1.88), **66** (PCE = 1.44) and **96** (PCE = 1.36) with the absence of F08[O–O] feature. This descriptor signifies the oxygen in the enamine of the Carbazole moiety and the anchoring functional groups (such as carboxylate, alkoxysilanes *etc.*) which may strengthen the  $\pi$ – $\pi$  interactions of the dye system to help in an increase in PCE [50].

The 2D atom pair descriptor F04[C–N] states that the frequency of carbon and nitrogen atoms at the topological distance 4 affects PCE positively. Thus, presence of this fragment in the dye may increase the PCE property as reported in dyes **50** (F04[C–N] = 22; PCE = 7.52), **101** (F04[C–N] = 17; PCE = 8.09) and **130** (F04[C–N] = 15; PCE = 9.8) and *vice versa* in case of **97** (PCE = 0.0538), **98** (PCE = 0.0387) in absence of F04[C–N] fragment. The presence of carbon and nitrogen atoms in the carbazole moiety may enrich the electron donating capability in the dyes in DSSCs which may increase the PCE values [54].

The ring descriptor nR10 denotes the number of 10-membered rings in the carbazole dyes with a positive contribution to the PCE. The presence of 10 membered ring N-annulated indenoperylene (electron donor) in the photo-chemically inactive segments of the carbazole dye can be conjugated *via* triple bond with an electron-acceptor for a metal-free donor or acceptor dye without the use of any co-adsorbate, which might be responsible for high PCE value of DSSCs [70]. Presence of 10

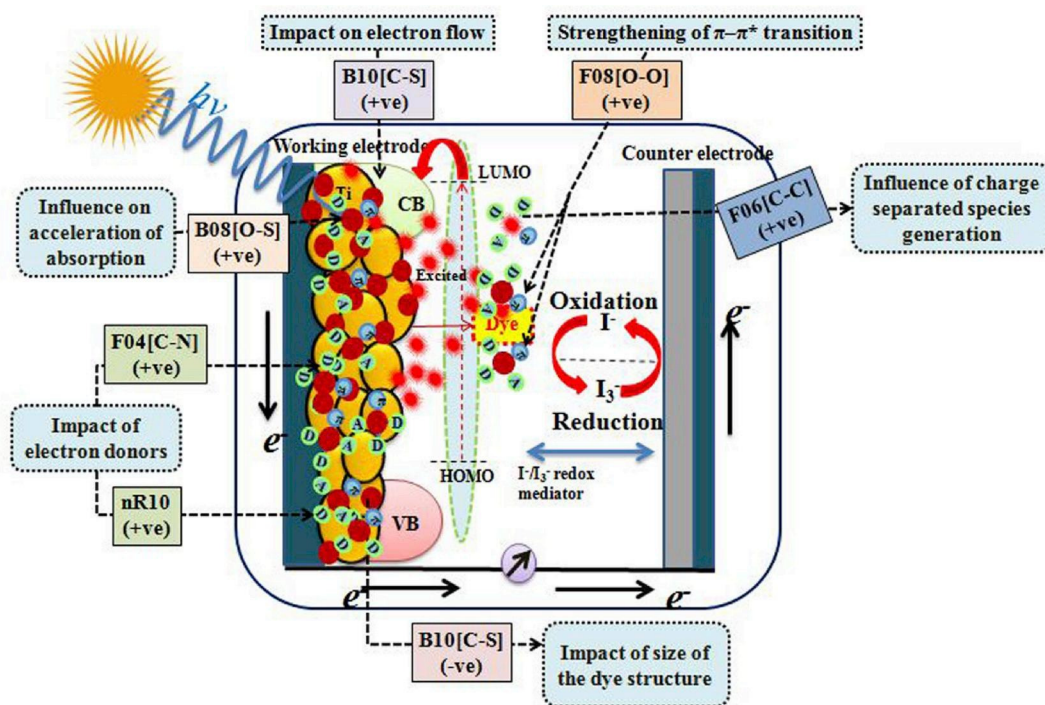


Fig. 10. Contribution of different descriptors on controlling PCE values of the carbazole dyes.

membered ring in the dyes correlate to higher PCE as observed in dyes **103** (nR10 = 1; PCE = 7.54), **130** (nR10 = 6; PCE = 9.8) and **131** (nR10 = 6; PCE = 7.6).

Another 2D atom pair descriptor B08[O-S] defines presence or absence of oxygen and sulfur atoms at the topological distance 8 bearing a positive effect to PCE. From the analysis of structures, we have found that these two atoms are present in methoxy and thiophene moieties of dyes where oxygen atom acts as linker for donor or acceptor whereas sulfur in thiophene moiety accelerates absorption by inducing bathochromic shift [47]. Thus, presence of such fragment in dye molecules may enhance the PCE property of dyes in DSSC as shown in dyes **50** (PCE = 7.52), **94** (PCE = 7.33) and **178** (PCE = 7.43) (descriptor value is 1 in all cases), while the absence of such fragments may detrimental to the PCE property as shown in compounds **91** (PCE = 0.19), **157** (PCE = 0.21) and **159** (PCE = 0.31).

The negative regression coefficient of 2D atom pair descriptors such as B04[O-S] (which means presence or absence of oxygen and sulfur atoms at topological distance 4) and F06[O-S] (indicates frequency of oxygen atom and sulfur at topological distance 6) indicate that presence of these fragments may reduce the PCE property of dyes in DSSC as evidenced by the dyes **12** (PCE = 1.79), **13** (PCE = 0.55) and **92** (PCE = 0.24) for B04[O-S] descriptor where the value of descriptor is 1 in all cases; **66** (PCE = 1.44), **154** (PCE = 0.07) and **156** (PCE = 0.06) for F06[O-S] descriptor where the numerical value is 1 in all cases and *vice versa* in case of dyes **94** (PCE = 7.33), **101** (PCE = 8.09) and **133** (PCE = 9.32), where both features are absent. These fragments may cause weak interaction between the semiconductor and carbazole dyes due to steric hindrance which make difficulty in the electron flow to the semiconductor.

The 2D atom pair descriptor B10[C-S] states the presence or absence of carbon and sulfur atoms at topological distance 10 offers a positive effect on the PCE property. Presence of this fragment favors the bulkiness of the dyes which may enhance the sensitized wide bandgap in the nano-structured photoelectrode [53]. As a result, the PCE values may increase in the presence of this fragment as shown in the dyes **50** (PCE = 7.52), **103** (PCE = 7.54) and **178** (PCE = 7.43). On the other hand, the absence of this feature may decrease the PCE property as shown in dyes

**32** (PCE = 0.695), **53** (PCE = 0.99) and **112** (PCE = 0.96).

The 2D atom pair descriptors B06[N-S] (presence or absence of nitrogen and sulfur atoms at the topological distance 6), B04[N-O] (presence or absence of nitrogen and oxygen atoms at topological distance 4), F06[N-O] (frequency of nitrogen and oxygen atoms at topological distance 6) and B02[C-S] (presence or absence of carbon and sulfur atoms at topological distance 2) contribute positively towards the PCE property of dyes in DSSC. The positive regression coefficients of these parameters suggest that the presence of such fragments in the Carbazole dyes enhances the PCE property as shown in the dyes **94** (PCE = 7.33), **99** (PCE = 7.58), **100** (PCE = 6.98) (for B06[N-S] descriptor, the descriptor value is "1" in all cases), **50** (PCE = 7.52), **103** (7.54), **178** (7.43) (for B04[N-O] descriptor, the numerical value of the descriptor is "1" in all cases), **104** (PCE = 6.93), **134** (PCE = 6.16), **135** (PCE = 6.33) (for F06[N-O], the descriptor value is "1" in all cases), **130** (PCE = 9.8), **131** (PCE = 7.6) (for B02[C-S], descriptor value is "1" in all cases). On the other hand, the dyes **124** (PCE = 1.04), **126** (PCE = 0.91) (the numerical value of these descriptors is "0" in all cases) showed lower PCE values for the Carbazole dyes due to absence of such fragments.

The negative regression coefficients of the 2D atom pair descriptor F06[N-N] (indicating the frequency of 2 nitrogen atoms at the topological distance 6), the atom type E-state descriptor NaasC (representing the number of atoms of aasC (-C(-)-)) and the constitutional descriptor N% (indicates the percentage of nitrogen atoms) suggested that the presence of these specific features might reduce the PCE values as shown in case of dyes **138** (PCE = 2.17), **150** (PCE = 2.49) for F06[N-N], where the descriptor values are 2 and 1, respectively; **62** (PCE = 1.88), **116** (PCE = 1.78) for NaasC, where the descriptor values are 15 and 18, respectively; **91** (PCE = 0.89), **176** (PCE = 0.07) for N% with values of 6.5 and 6.2, respectively, and *vice versa* in case of dyes **132** (PCE = 12.5) and **133** (PCE = 9.32), where F06[N-N] fragment is absent; **99** (PCE = 7.58) and **101** (PCE = 8.09), where NaasC descriptor value is 9 for both dyes; **133** (PCE = 9.32) and **99** (PCE = 7.58) for which N% values are 2 and 1.8, respectively.

The mechanistic interpretation of the PCE property of carbazole dyes from all models is schematically portrayed in Fig. 10. The scatter plots of observed vs. predicted PCE property related to the carbazole dyes for all

**Box 7**

IM1

$$\text{PCE} = 6.155 + 0.278 \times \text{F08}[\text{C} - \text{N}] + 3.681 \times \text{nPyrimidines} - 0.661 \times \text{F01}[\text{C} - \text{N}] - 0.170 \times \text{StsC}$$

IM2

$$\text{PCE} = 5.567 + 0.294 \times \text{F08}[\text{C} - \text{N}] + 3.354 \times \text{nPyrimidines} - 0.622 \times \text{F01}[\text{C} - \text{N}] - 0.143 \times \text{nCsp}$$

IM3

$$\text{PCE} = 2.857 - 2.974 \times \text{B08}[\text{N} - \text{N}] - 1.99 \times (\text{C} - 041) + 0.498 \times \text{nHAcc} - 0.267 \times \text{StsC}$$

IM4

$$\text{PCE} = 4.471 - 3.186 \times \text{B08}[\text{N} - \text{N}] + 2.919 \times \text{nPyrimidines} - 1.278 \times \text{F04}[\text{N} - \text{S}] - 1.016 \times \text{nR\#C} -$$

IM5

$$\text{PCE} = 5.714 + 0.241 \times \text{F08}[\text{C} - \text{N}] - 0.068 \times \text{ETA\_dBeta} - 0.634 \times \text{F01}[\text{C} - \text{N}] + 3.729 \times \text{nPyrimidines}$$

the PLS models are depicted in Fig. S12 in Supplementary material.

### 3.7. Dataset 7: modeling of the PCE property of diphenylamine dyes

The identified significant descriptors obtained from the five PLS models are illustrated in Box 7 with their respective contributions to the PCE property. The mechanistic interpretation of all the descriptors provided below with reasonable examples.

The 2D atom pair descriptor F08[C-N] defines the frequency of carbon and nitrogen atoms at the topological distance 8 which contribute positively towards the PCE. This fragment is a part of many electron donating groups (EDGs) in this dye system (seen in more than 50 structures in the form of hexyloxy amines and ethylhexyloxy amino groups, dithiadiazole groups) [71]. Therefore, presence of this fragment in the dyes may enhance the PCE as observed in the dyes 15 (F08[C-N] = 16; PCE = 6.66), 26 (F08[C-N] = 16; PCE = 7.1) and 27 (F08[C-N] = 20; PCE = 8) and *vice versa* in dyes 14 (F08[C-N] = 2; PCE = 3), 33 (F08[C-N] = 1; PCE = 0.44) and 35 (F08[C-N] = 1; PCE = 0.4).

The negative regression coefficient of B08[N-N] (representing presence/absence of the two nitrogen atoms at the topological distance

8) and F01[C-N] (designating the frequency of carbon and nitrogen atoms at topological distance 1) indicate that presence of these fragments in the dyes may reduce the PCE of DSSCs as shown in dyes 29 (B08[N-N] = 1 & F01[C-N] = 8; PCE = 1.99), 34 (B08[N-N] = 1 & F01[C-N] = 10; PCE = 1) and 35 (B08[N-N] = 1 & F01[C-N] = 11; PCE = 0.44). In contrast, the dyes having no such fragments may have enhanced PCE property as shown in dyes 15 (B08[N-N] = 0 & F01[C-N] = 6; PCE = 6.66), 17 (B08[N-N] = 0 & F01[C-N] = 6; PCE = 6.19) and 26 (B08[N-N] = 0 & F01[C-N] = 6; PCE = 7.05). Due to the presence of these groups, there is an effect on the polarity of the dye molecules which may cause dye aggregation on semiconductors mesoporous layer followed by minimization of photon absorption [72].

The functional group count descriptor nPyrimidines represents the number of pyrimidines present in the structure of a dye. The positive regression coefficient of the descriptor indicates the presence of pyrimidine ring favors the PCE property as shown in dyes 7 (nPyrimidines = 1; PCE = 7.05) and 8 (nPyrimidines = 1; PCE = 7.64), while dyes 14 (nPyrimidines = 0; PCE = 3), 28 (nPyrimidines = 0; PCE = 2.8) and 34 (nPyrimidines = 0; PCE = 1) show lower range of PCE values due to absence of the pyrimidine ring. Due to presence of pyrimidine in the dye,

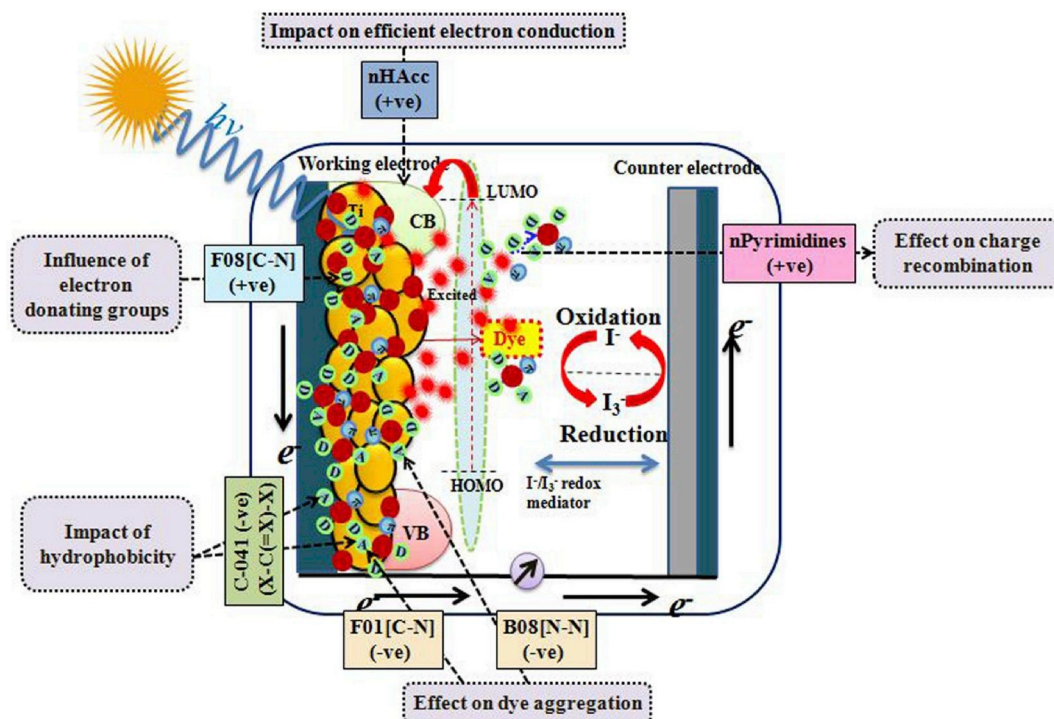


Fig. 11. Contribution of different descriptors on controlling PCE values of the Diphenylamine dyes.



the LUMOs lay over the conduction band edge of TiO<sub>2</sub>, while their HOMOs are under the reduction potential energy of the electrolytes (I<sup>−</sup>/I<sub>3</sub><sup>−</sup>); thus, the ability of electron transfer from the dye (excited state) to TiO<sub>2</sub> mesoporous layer increases, and charge regeneration occurs effectively. Thus, the effective charge regeneration may enhance the PCE property of dye molecules in DSSC [73].

The atom centered fragment C-041 [48,49] corresponds to X-C (=X)-X, where, X can be any electro negative atom O, N, S, P, Se and halogens connected with the carbon atom; it has a negative impact on the PCE property of dyes as found in dyes **33** (C-041 = 1; PCE = 0.44), **34** (C-041 = 2; PCE = 0.4) and **35** (C-041 = 2; PCE = 1). On the other hand, the dyes with lower numerical values of the descriptor show higher PCE values as observed in case of dyes **3** (C-041 = 0; PCE = 5.4), **7** (C-041 = 0; PCE = 7.05) and **27** (C-041 = 0; PCE = 8). This fragment favours the hydrophobicity of dyes by acting as a buffer between semiconductor and the electrolyte which leads to an effect on charge recombination followed by reduction of PCE property.

Another functional group count descriptor nHAcc stands for the number of acceptor atoms for H-bonds (N, O and F) which contributes positively to the PCE property of diphenylamine dyes. Thus, the presence of higher number of acceptor atoms in the dyes for hydrogen bonding favors forming a closely packed structure that increases its conductivity [74] as well as the PCE values as shown in the dyes **27** (nHAcc = 10; PCE = 8), **8** (nHAcc = 8; PCE = 7.64) and **7** (nHAcc = 8; PCE = 7.05). In contrast, the PCE property decreases when there is a decrease in the numerical value of this descriptor as shown in case of dyes **10** (nHAcc = 4; PCE = 1.99) and **33** (nHAcc = 5; PCE = 0.44).

The atom type E-state descriptor StsC (Sum of tsC E-states ≡C—), the constitutional descriptor nCsp (number of sp hybridized Carbon atoms) and the extended topochemical atom descriptor, ETA\_dBeta (measuring the relative unsaturation content (Δβ)) contribute negatively towards the PCE property of diphenylamine dyes as indicated by negative regression coefficient of these descriptors. This means that with an increase in the numerical values of the mentioned descriptors, the PCE value decreases as observed in case of dyes **10** (StsC = 8.29 & nCsp = 3, ETA\_dBeta = 7; PCE = 1.99), **14** (StsC = 8.31 & nCsp = 3, ETA\_dBeta = 8; PCE = 3) and *vice versa* in case of dyes **7** (StsC = 1.64 & nCsp = 1, ETA\_dBeta = 1.5; PCE = 7.05), **8** (StsC = 1.67 & nCsp = 1, ETA\_dBeta = −5; PCE = 7.64).

The mechanistic interpretation of the diphenylamine dyes from all models is schematically portrayed in Fig. 11. The scatter plots of observed vs. predicted PCE property related to the Phenothiazine dyes for all the PLS models are depicted in Fig. S13 in Supplementary material.

We have also checked the applicability domain of all the individual models (IM1-IM5) developed from the seven datasets using the DModX approach. Based on the AD study, we have found that all the training set dyes are within the stipulated D-critical value under 99% confidence limit for the Triphenylamine (1.7622–1.879), Phenothiazine (1.768), Indoline (1.598–1.916), Porphyrin (1.584–1.853), Coumarin (2.536–2.912), Carbazole (1.638–1.904) and Diphenylamine (2.135–3.911) datasets, respectively. In case of test set dyes, **132** (for model IM-1), **167** (for model IM-2), **165** (for model IM-3) and **160** (for model IM-5) for Triphenylamine dataset (Figs. S14–S18); **181**, **28** and **160** (IM1-IM5) for Phenothiazine dataset (Figs. S19–S23); **151** (IM1-IM5) for Indoline dataset (Figs. S24–S28); **270** (IM1-IM5) for Porphyrin dataset (Figs. S29–S33); **129** (IM1-IM-5) for Carbazole dataset (Figs. S34–S38) are identified as out of AD. In case of Coumarin (Figs. S39–S43) and Diphenylamine (Figs. S44–S48) datasets, all the test set compounds are within the AD zone. Considering the huge number of dyes, almost 99% of dyes pass the AD test, and their predictions are completely reliable.

We have performed Y-scrambling or Y-randomization analysis to check whether we got our model accidentally. But interestingly all scrambled models failed to achieve R<sup>2</sup> and Q<sup>2</sup> values of more than or equal to 0.5 which suggest that our developed models had not generated

by chance. The quality of these randomized models is so bad that no single R<sup>2</sup> value is over 0.1 and all Q<sup>2</sup> values are in negative range. The complete results can be found in Table S1 and the figures can be found in Figs. S49–S55 in the Supplementary materials file.

### 3.8. Overview of the obtained interpretation from QSPR models

We have developed multiple QSPR models for around 1200 organic dyes classified under seven chemical classes. The developed individual and consensus models helped us to identify the essential structural fragments and physicochemical features of the studied dyes which are responsible for variation of the PCE values in DSSCs. The models offer a series of mechanistic interpretation of the variation of PCE values with molecular structures of a large number of dyes which can be employed by the investigators to reduce the experimental testing, time, and money by several folds. Moreover, the exploratory information may help to design new, improved lead dyes for all seven classes. The quantitative structural analysis lead to the following interpretations for efficient PCE of individual chemical classes:

#### 3.8.1. Triphenylamine dyes

The fragment =N- lowers the tendency of localized π-π\* transition due to ICT transition from the triphenylamine donor to the anchor group and lowers the absorption maxima followed by a decrease of PCE values. The fragments Al-C(=X)-Al (Al: aliphatic, X: O, N, S, P, Se, halogens) and X-CR..X contribute to the hydrophobicity and act as a buffer between the semiconductor and the electrolyte which prevent the back-transfer of electrons from the semiconductor's conduction band to the redox couple and ultimately lower the PCE. Presence of imides (thio) favours dye hydrolysis which helps the aggregation of the dye over the TiO<sub>2</sub> surface and improves the recombination reaction between redox electrolyte and electrons in the TiO<sub>2</sub> nanolayer. On the other hand, presence of O and S atoms in the donor groups narrows the absorption range of dyes which causes latency decrease of rapid π-conjugation. Thus, all these fragments need to be avoided during the design of triphenylamine dyes.

On the contrary, long branching in thiophene ring results in a slight hypsochromic and hypochromic effect in the ICT band where the steric hindrance is induced by the branched-chain increases the torsion angle between the triphenylamine moiety and the thiophene unit. This torsion impedes good delocalization of the π electrons and blue-shifts the position of the ICT band and augment the absorption. Aliphatic imines (nRC = N) also help in the mobilization of free electrons to offer better PCE.

#### 3.8.2. Phenothiazine dyes

H attached to C0(sp<sup>3</sup>) with 1X attached to next C fragment favors the lipophilicity or hydrophobicity of the dyes which causes alterations in the energy cascade as a result of the physicochemical alterations such as poor solubility on semiconductors porous layer decreases the PCE values. Again, the mean first ionization potential on a scaled C atom is related to polarity of the molecules due to presence of small electro-negative atoms which results in the aggregation of dyes on the semiconductor and ultimately lowers the PCE.

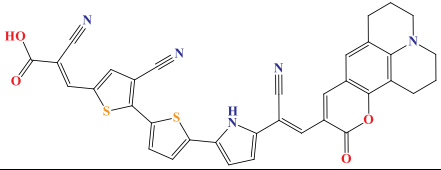
Higher number of O atoms is good as they are involved in the conduction of electrons towards the excitation state due to their natural tendency to form a closely packed structure. As a result, the higher conductivity carriers in the dyes enhance the PCE. Fragment like #CR/R = C = R (= double bond; # a triple bond) is good for mobilization of free electrons and ultimately flow of current in the solar system. Again, large surface area of the dyes may affect the photon capturing ability due to the sensitized wide band gap in the photo electrode which is one of the reasons for high PCE values. The presence of C and O atoms at topological distance 8 signifies the effect of donor and an additional donor through a linkage in the dye system which helps to achieve absorption band broadening followed by an increase in the PCE. The N and O atoms

**Table 2**  
Calculated descriptors and predicted % PCE of the designed coumarin dyes (NCM1 to NCM10).

Dye	Structure	Computed Descriptor					Predicted (%) PCE
		nR = Ct	C-034	T(S..S)	nR#C-	C-040	
NCM1		3	6	0	0	4	8.93
NCM2		3	6	0	0	4	8.93
NCM3		4	6	0	0	5	10.62
NCM4		3	6	0	0	5	9.46
NCM5		4	6	0	0	5	10.62
NCM6		4	6	0	0	5	10.62
NCM7		3	6	3	0	5	9.17
NCM8		3	6	0	0	5	9.46
NCM9		4	6	3	0	5	10.32



(continued on next page)

Table 2 (continued)

Dye	Structure	Computed Descriptor					Predicted (%) PCE
		nR = Ct	C-034	T(S..S)	nR#C-	C-040	
NCM10		3	6	3	0	5	9.17

at the topological distance 4 signifies to a strong cyano acceptor and a chelating anchoring mode of the carboxylate ion which plays a crucial role to regulate the PCE property of dyes.

### 3.8.3. Indoline dyes

Dyes with nitro group,  and  fragments experience poor  $\pi$ - $\pi^*$  transition (less reactive than imines) which results in a slow energy cascade mechanism followed by lower PCE values. O-S and C-N atoms at the topological distance 10 represent bulkiness of dyes which may weaken the interactions between the semiconductor and the dye due to steric hindrance followed by restriction of the transfer of electrons from the dye to the semiconductor. Therefore, smaller fragments with lower branching is expected to increase PCE.

Ring quaternary carbon with  $sp^3$  hybridization and non-aromatic conjugated carbon with  $sp^2$  hybridization are good for PCE as this feature is essential for tunable absorption properties, and they produce high molar extinction coefficients leading to improved energy level reactions in the solar cell. Indoline dyes containing donor groups and groups with non-planar structure are very important for the PCE property. Fragments with O and S atoms at the topological distances 5 and 9 regulate the electron density delocalization which is favorable for the  $\pi$ -bond conjugation. As a result, the molar extinction coefficient of the dye is enhanced which leads to the bathochromic shift of the absorption spectrum followed by better PCE.

### 3.8.4. Porphyrin dyes

Presence of methine bridges ( $=CH-$ ), a non-polar group makes the negative shift in the solvatochromic properties of the dye which cannot adhere properly to the semiconductor which results in a negative effect on the absorption and stability of the dye. Additionally, fragments like  $RCO-N<$  or  $>N-X = X$  and 8 membered rings could be avoided for better PCE values.

Aliphatic and aromatic amines, the N-O at topological distances 3 and 10 fragments have a contribution to the special stability (conjugation) and allow a smaller HOMO-LUMO gap, followed by the red shift of the absorption spectrum. Thus, the conduction valence edge level increases, and as a result, the PCE values of DSSCs increases. Again, C and O at the topological distances 6 and 7 with heavy metal (X) suggest the long-chain alkoxy group which impair interfacial back electron transfer reaction help in electron donating ability, and the C in meso aryl substituted portion of Zinc porphyrin acts as a donor in the dye, and these may increase the PCE property. The average connectivity index of order 4 related to surface area of dyes is directly related to light-harvesting capability which could be achieved maximum when the surface area of the dyes is large.


### 3.8.5. Coumarin dyes

The topological distance between two sulfur atoms needs to be lower in case of coumarin dyes. One or two thiophene group(s) would be optimal, preferably one. Again, if two thiophene groups are present in a coumarin dye, they should be connected with a single bond to have the lowest topological distance between the S atoms. Absence of a non-terminal carbon with " $sp$ " hybridization is also good for enhancement

of PCE.

The aliphatic nitriles increase the electron withdrawing ability which is meant to interact with the electron-donating groups present in the dye. Thus, this kind of dipolar fragment provides directionality of the electronic orbitals in the excited state which leads to photo-induced charge-transfer excitation of the dye-to  $TiO_2$  bands. The non-aromatic conjugated  $sp^2$  hybridized C atoms and aliphatic tertiary carbon with " $sp^2$ " hybridization help in the conjugation extension units increase which helps in broadening the absorption and photosensitization properties of the coumarin dyes followed by better PCE. Fragments like  $R-CR..X$ ,  $R-C(=X)-X$  or  $R-C\#X$  or  $X = C = X$  ( $\#$ : triple bond;  $..$ : aromatic bond as in benzene or delocalized bonds such as the N-O bond in a nitro group;  $..$ : aromatic single bonds as the C-N bond in pyrrole) are good for PCE.

### 3.8.6. Carbazole dyes

Fragments with O and S atoms at the topological distances 4 and 6 may cause weak interaction between the semiconductor and carbazole dyes due to steric hindrance which causes difficulty in the electron flow to the semiconductor. Presence of  fragment and 2 N atoms at the topological distance 6 results in poor  $\pi$ - $\pi^*$  transition followed by lower PCE values.

In contrast, 10-membered ring N-annulated indenoperylene (electron donor) in the photo-chemically inactive segments of the carbazole dye can be conjugated via triple bond with an electron-acceptor for a metal-free donor or acceptor dye without use of any co-adsorbate which might be responsible for high PCE values. Fragments with two O atoms at the topological distance 8 signify the oxygen in the enamine of the carbazole moiety and the anchoring functional groups (such as carboxylate, alkoxy silanes etc.), which may strengthen the  $\pi$ - $\pi$  interactions of the dye system to help in an increase of PCE values. Again, a fragment with C and N atoms at the topological distance 4 moiety may enrich the electron donating capability in the dyes which may increase the PCE values. The fragment with the O and S atoms at the topological distance 8 present in methoxy and thiophene moieties of dyes where the O atom acts as linker for donor or acceptor and S in thiophene moiety accelerates absorption by inducing bathochromic shift.

### 3.8.7. Diphenylamine dyes

Fragments like  $X-C(=X)-X$  and  $\equiv C$ —favour the hydrophobicity of dyes by acting as a buffer between semiconductor and the electrolyte which have an effect on charge recombination followed by reduction of PCE property. Again, number of  $sp$  hybridized C atoms and higher number of unsaturation content in the dye result in the reduction of PCE values. Again, fragments like 2 N atoms at the topological distance 8 and C and N atoms at the topological distance 1 have an effect on the polarity of the dye which may cause dye aggregation on semiconductor's mesoporous layer followed by minimization of photon absorption.

The pyrimidine scaffold has the ability of electron transfer from the dye (excited state) to  $TiO_2$  mesoporous layer, and charge regeneration is done effectively for higher PCE. Again, presence of more acceptor atoms for H-bonds (N, O and F) in the dyes for hydrogen bonding favour to form a closely packed structure that increases its conductivity followed



by PCE. The electron donating groups like hexyloxy amines, ethyl-hexyloxy amino groups and dithiadiazole groups help in an increase of PCE values.

#### 4. Design of new dyes

The success of a QSPR model is in its implementation to design new dyes with an enhanced response (here %PCE). Therefore, we have tried to design power conversion efficient dyes based on our generated best models. In our previous studies, we have already designed indoline, diphenylamine and tetrahydroquinoline dyes [17,24,25] employing QSPR analysis and quantum chemical studies. In the present study, we have considered the coumarin chemical class for designing purpose due to its least %PCE value compared to all other studied chemical classes. Among the studied 58 coumarin dyes, the highest experimental %PCE value is 7.4 and only 5 dyes has %PCE value more than 6. So, undoubtedly coumarin dye is the least explored and developed classes of dyes among the studied ones. Thus, model IM5 is considered for the design of coumarin dyes as it is the best developed model as discussed in the Results and Discussion section. Based on the interpretation of the modeled descriptors of equation IM5 of Box 5, we have designed 10 coumarin dyes (NCM1-NCM10) (See Table 2) for which theoretical predictions showed % PCE ranging from 8.93 to 10.62. Compared to the highest reported experimental %PCE of 7.4 in the studied dataset, the designed dyes showed a 20.68–43.51% increase in PCE values. In order to check the AD of the designed dyes, we have applied the DModX approach under 99% confidence limit and found that all 10 compounds reside under the modeled D-critical value of 2.912. The AD plot of the designed coumarin dyes is illustrated in Fig. S56 in Supplementary materials.

#### 5. Conclusion

In the overall conclusion, the identified requisite fragments are important for photophysical properties and optimal balance of short-circuit current (JSC) and the open-current voltage (VOC). The present manuscript has largely explored the 2D structural fragments features, as the quantum and electrochemical analyses for large number of dyes are time consuming. Aliphatic or aromatic amines/imines, thiophenes, pyrimidines, and pyrrole ring systems with an overall requirement in terms of  $\pi$  electrons or aromatic system for the mobilization of free electrons in the form of current, more acceptor atoms for H-bonds in the form of D–A– $\pi$ –A structure framework, polarity, optimum chain length to avoid hydrophobicity of dyes are the overall features for most of the chemical classes with some specific exceptions to individual chemical classes. The interpreted features from the QSPR models helped us in designing of power conversion efficient ten coumarin dyes. The potentially best designed dye showed predicted %PCE value of 10.62 which is a 43.5% increase compared to the existing coumarin dye with the highest PCE value in the modeled dataset. Our suggested exploratory features of dyes from other chemical classes may also help to design more efficient dyes similar to the case of coumarins saving time and money.

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

JGK thanks the Ministry of Chemicals & Fertilizers, Department of Pharmaceuticals, Government of India and the National Institute of Pharmaceutical Education and Research Kolkata (NIPER-Kolkata) for providing financial assistance in the form of a fellowship. PKO thanks

the UGC, New Delhi for financial assistance in the form of a fellowship. SK and JL thankful to the Department of Energy (grant number: DE-SC0018322) and the NSF EPSCoR (grant number: OIA-1757220) for financial support. KR thanks CSIR, New Delhi for financial assistance under a Major Research project (CSIR Project No. 01(2895)/17/EMR-II).

#### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.nanoen.2020.104537>.

#### References

- [1] M. Hagfeldt, A. Grätzel, Molecular photovoltaics, *Acc. Chem. Res.* 33 (2000) 269–277.
- [2] M. Pastore, S. Fantacci, F. De Angelis, Modeling excited states and alignment of energy levels in dye-sensitized solar cells: successes, failures, and challenges, *J. Phys. Chem. Chem.* 117 (2013) 3685–3700.
- [3] A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo, H. Pettersson, Dye-sensitized solar cells, *Chem. Rev.* 110 (2010) 6595–6663.
- [4] K. Sharma, V. Sharma, S.S. Sharma, Dye-sensitized solar cells: fundamentals and current status, *Nanoscale Res. Lett.* 13 (2018) 381, <https://doi.org/10.1186/s11671-018-2760-6>.
- [5] A. Baheti, K.R. Justin Thomas, C.T. Li, C.P. Lee, K.C. Ho, Fluorene-based sensitizers with a phenothiazine donor: effect of mode of donor tethering on the performance of dye-sensitized solar cells, *ACS Appl. Mater. Interfaces* 7 (2015) 2249–2262.
- [6] W. Zhang, Y. Wu, H. Zhu, Q. Chai, J. Liu, H. Li, X. Song, W.H. Zhu, Rational molecular engineering of indoline-based D-A- $\pi$ -A organic sensitizers for long-wavelength-responsive dye-sensitized solar cells, *ACS Appl. Mater. Interfaces* 7 (2015) 26802–26810.
- [7] A. Mishra, M.K.R. Fischer, P. Büuerle, Metal-Free organic dyes for dye-Sensitized solar cells: from structure: property relationships to design rules, *Angew. Chem. Int. Ed.* 48 (2009) 2474–2499.
- [8] M. Liang, J. Chen, Arylamine organic dyes for dye-sensitized solar cells, *Chem. Soc. Rev.* 42 (2013) 3453–3488, 2013.
- [9] R.S. Ashraf, I. Meager, M. Nikolka, M. Kirkus, M. Planells, B.C. Schroeder, S. Holliday, M. Hurhangee, C.B. Nielsen, H. Sirringhaus, I. McCulloch, Chalcogenophene comonomer comparison in small band gap diketopyrrolopyrrole-based conjugated polymers for high-performing field-effect transistors and organic solar cells, *J. Am. Chem. Soc.* 137 (2015) 1314–1321.
- [10] U.B. Cappel, M.H. Karlsson, N.G. Pschirer, F. Eickemeyer, J. Schöneboom, P. Erk, G. Boschloo, A. Hagfeldt, A broadly absorbing perylene dye for solid-state dye-sensitized solar cells, *J. Phys. Chem. C* 113 (2009) 14595–14597.
- [11] N.J. Cherepy, G.P. Smestad, M. Grätzel, J.Z. Zhang, Ultrafast electron injection: implications for a photoelectrochemical cell utilizing an anthocyanin dye-sensitized TiO<sub>2</sub> nanocrystalline electrode, *J. Phys. Chem. B* 101 (1997) 9342–9351.
- [12] M.K. Nazeeruddin, F. De Angelis, S. Fantacci, A. Selloni, G. Viscardi, P. Liska, S. Ito, B. Takeru, M. Grätzel, Combined experimental and DFT-TDDFT computational study of photoelectrochemical cell ruthenium sensitizers, *J. Am. Chem. Soc.* 127 (2005) 16835–16847.
- [13] I. Choi, M. Ju, S. Kang, M. Kang, B. You, J. Hong, H.K. Kim, Structural effect of carbazole-based coadsorbents on the photovoltaic performance of organic dye-sensitized solar cells, *J. Mater. Chem. A* 32 (2013) 9114–9121.
- [14] L. Zhang, X. Yang, W. Wang, G.G. Gurzadyan, J. Li, X. Li, J. An, Z. Yu, H. Wang, B. Cai, A. Hagfeldt, L. Sun, 13.6% Efficient organic dye-sensitized solar cells by minimizing energy losses of the excited state, *ACS Energy Lett.* 4 (2019) 943–951.
- [15] S. Mathew, A. Yella, P. Gao, R. Humphry-Baker, B.F. E. Curchod, N. Ashari-Astani, I. Tavernelli, U. Rothlisberger, M. Khaja Nazeeruddin, M. Grätzel, Dye-sensitized solar cells with 13% efficiency achieved through the molecular engineering of porphyrin sensitizers, *Nat. Chem.* 6 (2014) 242.
- [16] K. Roy, S. Kar, R.N. Das, Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences, Academic Press (Elsevier), 2015.
- [17] S. Kar, J.K. Roy, J. Leszczynski, In silico designing of power conversion efficient organic lead dyes for solar cells using today's innovative approaches to assure renewable energy for future, *Npj Comput. Mater.* 3 (2017).
- [18] S. Kar, N. Sizochenko, L. Ahmed, V.S. Batista, J. Leszczynski, Quantitative structure-property relationship model leading to virtual screening of fullerene derivatives: exploring structural attributes critical for photoconversion efficiency of polymer solar cell acceptors, *Nano Energy* 26 (2016) 677–691.
- [19] J.K. Roy, S. Kar, J. Leszczynski, Optoelectronic properties of C60 and C70 fullerene derivatives: designing and evaluating novel candidates for efficient P3HT polymer solar cells, *Materials (Basel)* 12 (2019) 2282.
- [20] V. Venkatraman, B.K. Alsberg, A quantitative structure-property relationship study of the photovoltaic performance of phenothiazine dyes, *Dyes Pigments* 114 (2015) 69–77.
- [21] V. Venkatraman, M. Foscatto, V.R. Jensen, B.K. Alsberg, Evolutionary de novo design of phenothiazine derivatives for dye-sensitized solar cells, *J. Mater. Chem. A* 3 (2015) 9851–9860.
- [22] H. Li, Z. Zhong, L. Li, R. Gao, J. Cui, T. Gao, L.H. Hu, Y. Lu, Z.M. Su, H. Li, A cascaded QSAR model for efficient prediction of overall power conversion

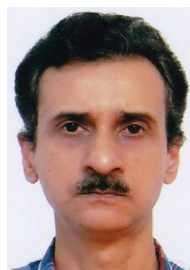
- efficiency of all-organic dye-sensitized solar cells, *J. Comput. Chem.* 36 (2015) 1036–1046.
- [23] S. Kar, J.K. Roy, D. Leszczynski, Power conversion efficiency of arylamine organic dyes for dye-sensitized solar cells (DSSCs) explicit to cobalt electrolyte: understanding the structural attributes using a direct QSPR approach, *Computation* 5 (2017) 2.
- [24] J.K. Roy, S. Kar, J. Leszczynski, Electronic structure and optical properties of designed photo-efficient indoline-based dye-sensitizers with D–A– $\pi$ –A framework, *J. Phys. Chem. C* 123 (2019) 6.
- [25] J.K. Roy, S. Kar, J. Leszczynski, Insight into the optoelectronic properties of designed solar cells efficient tetrahydroquinoline dye-sensitizers on TiO<sub>2</sub>(101) surface: first principles approach, *Sci. Rep.* 8 (2018) 10997.
- [26] V. Venkatraman, R. Raju, S.P. Oikonomopoulos, B.K. Alsberg, The dye-sensitized solar cell database, *J. Cheminf.* 10 (2018) 18.
- [27] H. Moriawaki, Y.S. Tian, N. Kawashita, T. Takagi, Mordred: a molecular descriptor calculator, *J. Cheminf.* 10 (2018) 4.
- [28] MarvinSketch software. <https://www.chemaxon.com>.
- [29] Dragon Version 7, 2016. <http://www.taletel.mi.it/index.htm>.
- [30] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, *J. Comput. Chem.* 32 (2011) 1466–1474.
- [31] [http://teqip.jdvu.ac.in/QSAR\\_Tools/DTCLab](http://teqip.jdvu.ac.in/QSAR_Tools/DTCLab).
- [32] P.K. Ojha, K. Roy, Comparative QSARs for antimalarial endochins: importance of descriptor-thinning and noise reduction prior to feature selection, *Chemometr. Intell. Lab. Syst.* 109 (2011) 146–161.
- [33] S. Das, P.K. Ojha, K. Roy, Multilayered variable selection in QSPR: a case study of modeling melting point of bromide ionic liquids No Title, *Int. J. Quant. Struct. Relat.* 2 (2017) 106–124.
- [34] K. Khan, E. Benfenati, K. Roy, Consensus QSAR modeling of toxicity of pharmaceuticals to different aquatic organisms: ranking and prioritization of the DrugBank database compounds, *Ecotoxicol. Environ. Saf.* 168 (2019) 287–297.
- [35] J. Roy, S. Ghosh, P.K. Ojha, K. Roy, Predictive quantitative structure-property relationship (QSPR) modeling for adsorption of organic pollutants by carbon nanotubes (CNTs), *Environ. Sci. Nano.* 6 (2019) 224–247.
- [36] S. Ghosh, P.K. Ojha, K. Roy, Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs, *Chemosphere* (2019) 545–555.
- [37] K. Roy, Quantitative structure-activity relationships (QSARs): a few validation methods and software tools developed at the DTC laboratory, *J. Indian Chem. Soc.* 95 (2018) 1497–1502.
- [38] K. Roy, R.N. Das, P. Ambure, R.B. Aher, Be aware of error measures. Further studies on validation of predictive QSAR models, *Chemometr. Intell. Lab. Syst.* 152 (2016) 18–33.
- [39] <https://www.minitab.com/en-us/products/minitab/>.
- [40] SIMCA-P, UMETRICS, Sweden, 2002. <https://umetrics.com/>.
- [41] K. Roy, P. Ambure, S. Kar, P.K. Ojha, Is it possible to improve the quality of predictions from an “intelligent” use of multiple QSAR/QSPR/QSTR models? *J. Chemom.* 32 (2018), e2992.
- [42] K. Roy, I. Mitra, On various metrics used for validation of predictive QSAR models with applications in virtual screening and focused library design, *Comb. Chem. High Throughput Screen.* 14 (2011) 450–474.
- [43] K. Roy, I. Mitra, P.K. Ojha, S. Kar, R.N. Das, H. Kabir, Introduction of  $r_m^2$ (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models, *Chemometr. Intell. Lab. Syst.* 118 (2012) 200–210.
- [44] C. Sakong, H.J. Kim, S.H. Kim, J.W. Namgoong, J.H. Park, J.H. Ryu, B. Kim, M. J. Ko, J.P. Kim, Synthesis and applications of new triphenylamine dyes with donor-donor-(bridge)-acceptor structure for organic dye-sensitized solar cells, *New J. Chem.* 36 (2012) 2025–2032.
- [45] A. Mahmood, Triphenylamine based dyes for dye sensitized solar cells: a review, *Sol. Energy* 123 (2016) 127–144.
- [46] V. Tamilavan, N. Cho, C. Kim, J. Ko, M.H. Hyun, Synthesis of triphenylamine-based thiophene-(N-aryl)pyrrole-thiophene dyes for dye-sensitized solar cell applications, *Tetrahedron* 68 (2012) 5890–5897.
- [47] M. Xu, S. Wenger, H. Bala, D. Shi, R. Li, Y. Zhou, S.M. Zakeeruddin, M. Grätzel, P. Wang, Tuning the energy level of organic sensitizers for high-performance dye-sensitized solar cells, *J. Phys. Chem. C* 113 (2009) 2966–2973.
- [48] J.V. Knop, W.R. Muller, K. Szymanski, N. Trinajstić, Computer Generation of Certain Classes of Molecules, SKTH, Kemija u industriji, Zagreb, 1985.
- [49] A.K. Ghose, G.M. Crippen, Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity, *J. Comput. Chem.* 7 (1986) 565–577.
- [50] K. Sharma, V. Sharma, S.S. Sharma, Dye-sensitized solar cells: fundamentals and current status, *Nanoscale Res. Lett.* 13 (2018) 381.
- [51] N.R. Neale, N. Kopidakis, J. Van De Lagemaat, M. Gra, A.J. Frank, Effect of a Coadsorbent on the Performance of Dye-Sensitized TiO<sub>2</sub> Solar Cells: Shielding versus Band-Edge Movement, (No. NREL/CP-590-38978), National Renewable Energy Lab. (NREL), Golden, CO (United States), 2005.
- [52] K. Jasim, Dye sensitized solar cells-working principles, challenges and opportunities, *Sol. Cells-Dye-Sens. Dev.* (2011) 172–204.
- [53] K. Jasim, Natural dye sensitized solar cell based on nanocrystalline TiO<sub>2</sub>, *Sains Malays.* 41 (2012) 10116.
- [54] J.S. Luo, Z.Q. Wan, C.Y. Jia, Recent advances in phenothiazine-based dyes for dye-sensitized solar cells, *Chin. Chem. Lett.* 27 (2016) 1304–1318.
- [55] C. Bauer, G. Boschloo, E. Mukhtar, A. Hagfeldt, Interfacial electron-transfer dynamics in Ru(tcterpy)(NCS) 3-sensitized TiO<sub>2</sub> nanocrystalline solar cells, *J. Phys. Chem.* 106 (2002) 12693–12704.
- [56] L. Zhang, J.M. Cole, Anchoring Groups for Dye-Sensitized Solar Cells, 2015.
- [57] Z. Yang, C. Shao, D. Cao, Screening donor groups of organic dyes for dye-sensitized solar cells, *RSC Adv.* 5 (2015) 22892–22898.
- [58] A.H. Ahlha, F. Nurosyid, A. Supriyanto, The chemical bonds effect of Amaranthus hybridus L. and Dracaena Angustifolia on TiO<sub>2</sub> as photo-sensitizer for dye-sensitized solar Cells (DSSC), in: AIP Conf. Proc., American Institute of Physics Inc., 2017.
- [59] S. Wang, Y. Dong, C. He, Y. Gao, N. Jia, Z. Chen, W. Song, The role of sp<sup>2</sup>/sp<sup>3</sup> hybrid carbon regulation in the nonlinear optical properties of graphene oxide materials, *RSC Adv.* 7 (2017) 53643–53652.
- [60] L.L. Sun, T. Zhang, J. Wang, H. Li, L.K. Yan, Z.M. Su, Exploring the influence of electron donating/withdrawing groups on hexamolybdate-based derivatives for efficient p-type dye-sensitized solar cells (DSSCs), *RSC Adv.* 5 (2015) 39821–39827.
- [61] M. Ishida, S. Woo Park, D. Hwang, Y. Bean Koo, J.L. Sessler, D. Young Kim, D. Kim, Donor-substituted  $\beta$ -functionalized porphyrin dyes on hierarchically structured mesoporous TiO<sub>2</sub> spheres. Highly efficient dye-sensitized solar cells, *J. Phys. Chem. C* 115 (2011) 19343–19354.
- [62] A. Yella, H.W. Lee, H.N. Tsao, C. Yi, A.K. Chandiran, M.K. Nazeeruddin, E.W. G. Diau, C.Y. Yeh, S.M. Zakeeruddin, M. Grätzel, Porphyrin-sensitized solar cells with cobalt (II/III)-based redox electrolyte exceed 12 percent efficiency, *Science* 334 (80) (2011) 629–634.
- [63] S. Huber, N. Hutter, R. Jordan, Effect of end group polarity upon the lower critical solution temperature of poly(2-isopropyl-2-oxazoline), *Colloid Polym. Sci.* 286 (2008) 1653–1661.
- [64] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, Dragon software: an easy approach to molecular descriptor calculations 56 (2006) 237–248.
- [65] Y. Zhang, Z. Sun, H. Wang, Y. Wang, M. Liang, S. Xue, Nitrogen-doped graphene as a cathode material for dye-sensitized solar cells: effects of hydrothermal reaction and annealing on electrocatalytic performance, *RSC Adv.* 5 (2015) 10430–10439.
- [66] K. Zhu, N.R. Neale, A. Miedaner, A.J. Frank, Enhanced charge-collection efficiencies and light scattering in dye-sensitized solar cells using oriented TiO<sub>2</sub> nanotubes arrays, *Nano Lett.* 7 (2007) 69–74.
- [67] P.S. Kalsi, Spectroscopy of Organic Compounds, New Age International, 2007.
- [68] N. Koumura, Z.S. Wang, S. Mori, M. Miyashita, E. Suzuki, K. Hara, Alkyl-functionalized organic dyes for efficient molecular photovoltaics, *J. Am. Chem. Soc.* 128 (2006) 14256–14257.
- [69] M.K.R. Fischer, S. Wenger, M. Wang, A. Mishra, S.M. Zakeeruddin, M. Grätzel, P. Baurle, D- $\pi$ -A sensitizers for dye-sensitized solar cells: linear vs branched oligothiophenes, *Chem. Mater.* 22 (2010) 1836–1845.
- [70] Z. Yao, M. Zhang, H. Wu, L. Yang, R. Li, P. Wang, Donor/acceptor indenoperylene dye for highly efficient organic, *Dye-Sens. Solar Cells* 137 (2015) 3799–3802.
- [71] S.H. Kang, I.T. Choi, M.S. Kang, Y.K. Eom, M.J. Ju, J.Y. Hong, H.S. Kang, H.K. Kim, Novel D- $\pi$ -A structured porphyrin dyes with diphenylamine derived electron-donating substituents for highly efficient dye-sensitized solar cells, *J. Mater. Chem. A* 1 (2013) 3977–3982.
- [72] M.K. Hossain, M.F. Pervaz, M.N.H. Mia, A.A. Mortuza, M.S. Rahaman, M.R. Karim, J.M.M. Islam, F. Ahmed, M.A. Khan, Effect of dye extracting solvents and sensitization time on photovoltaic performance of natural dye sensitized solar cells, *Results Phys.* 7 (2017) 1516–1523.
- [73] E.V. Verbitskiy, E.M. Cheprakova, J.O. Subbotina, A.V. Schepochkin, P. A. Slepukhin, G.L. Rusinov, V.N. Charushin, O.N. Chupakhin, N.I. Makarova, A. V. Metelitsa, V.I. Minkin, Synthesis, spectral and electrochemical properties of pyrimidine-containing dyes as photosensitizers for dye-sensitized solar cells, *Dyes Pigments* 100 (2014) 201–214.
- [74] C.P. Lee, C.T. Li, K.C. Ho, Use of organic materials in dye-sensitized solar cells, *Mater. Today* 20 (2017) 267–283.



**Mr. Jillella Gopala Krishna** is a PhD student in National Institute of Pharmaceutical Education and Research (NIPER), Kolkata, India. At present, he is working in Drug Theoretics and Cheminformatics (DTC) lab in the Department of Pharmaceutical Technology, Jadavpur University, Kolkata, India. He pursued his Master's from NIPER, Mohali, India and Bachelors from SPSP, Tirupati, India. His work focuses specifically on the Quantitative Structure Activity Relationship, chemometric modeling and Machine learning studies on pharmaceuticals and industrial chemicals.



**Dr. Probir Kumar Ojha** is a post-doctoral fellow of the University Grant Commission (UGC), New Delhi, currently pursuing his research in Jadavpur University (JU), India. He completed his B. Pharmacy (2006) from JU and M. Pharmacy (2008) from BIT, Mesra, India. He has completed his PhD from JU (2014), India. He is a former visiting fellow in the University of Gdańsk, Poland under Marie-Curie research fellowship. Dr. Ojha is working in the field of QSAR and chemometric modeling for the span of 12 years. He has published 40 research and review articles in various reputed journals. Dr. Ojha is actively associated as a reviewer in various peer-reviewed journals.



**Dr. Kunal Roy** is a Professor in the Department of Pharmaceutical Technology, Jadavpur University, India. He has been a recipient of Commonwealth Academic Staff Fellowship (University of Manchester, 2007) and Marie Curie International Incoming Fellowship (University of Manchester, 2013). The field of his research interest is Quantitative Structure-Activity Relationship (QSAR) with application in Drug Design, Property Modeling and Predictive Ecotoxicology. Dr. Roy has published more than 300 research papers in refereed journals (current SCOPUS h index 42; total citations till date 8406). Dr. Roy is the Co-Editor-in-Chief of Molecular Diversity (Springer Nature) and Editor-in-Chief of IJQSPR (IGI Global).



**Dr. Supratik Kar** is a postdoctoral research associate in Jackson State University (JSU), Mississippi, USA. He has completed his B. Pharm. (2008) and M. Pharm. (2010) from JU, India securing first position in both degrees. He has earned his Ph.D. (2015) from JU, India. He is a former visiting researcher at the University of Gdańsk, Poland under Marie-Curie research fellowship. He has experience in QSAR and chemometric modeling studies for over ten years. He has published 61 research and review articles, 12 book chapters and 2 QSAR textbooks. He is actively associated as a reviewer for 43 peer-reviewed journals and performed 142 reviews.



**Prof. Jerzy Leszczynski** is a Professor of Chemistry and a President's Distinguished Fellow at JSU, Mississippi, USA. Dr. Leszczynski is a computational quantum chemist. He has published almost 1000 referred papers in leading journals and over 70 book chapters. He has edited and co-edited 42 books and has advised 30 students who already received their Ph.D. degrees. He has been cited almost 31,000 times and carries an H-index of 84 (Google Scholar). His international awards include: Guest Professorship, CAS, Shanghai, China 2002; an Honorary Doctorates, Dnipropetrovsk National University, Ukraine 2003 and, Wroclaw University of Technology, Poland 2016, the Maria Skłodowska-Curie's Medal Polish Chemical Society, 2007. He has been shortlisted for European Academy of Sciences 2002; International Academy of Engineering, 2003; Ukrainian Ecological Academy of Sciences, 2003 and European Academy of Sciences, Arts and Humanities, 2004.